This product is part of the RAND Corporation technical report series. Reports may include research findings on a specific topic that is limited in scope; present discussions of the methodology employed in research; provide literature reviews, survey instruments, modeling exercises, guidelines for practitioners and research professionals, and supporting documentation; or deliver preliminary findings. All RAND reports undergo rigorous peer review to ensure that they meet high standards for research quality and objectivity.

TECHNICAL REPORT

Enabling long-term access to scientific, technical and medical data collections

Jeff Rothenberg, Stijn Hoorens

Prepared for the British Library

RAND EUROPE

The research described in this report was prepared for and funded by the British Library.

Published 2010 by the RAND Corporation
1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1200 South Hayes Street, Arlington, VA 22202-5050
4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213-2665
Westbrook Centre, Milton Road, Cambridge CB4 1YG, United Kingdom
RAND URL: http://www.rand.org/
RAND Europe URL: http://www.rand.org/randeurope
To order RAND documents or to obtain additional information, contact
Distribution Services: Telephone: (310) 451-7002;
Fax: (310) 451-6915; Email: order@rand.org

Preface

The British Library (BL) is considering its potential role in the intake, curation, archiving and preservation of selected scientific, technical and medical (STM) reference datasets, with the aim of providing access to and manipulation of these datasets for research purposes. In order to develop an appropriate strategy, or a set of alternative tactics, to fulfil this mission, the BL requires an analysis of the characteristics and uses of such reference data collections. The BL commissioned RAND Europe to perform a scoping study that investigates its potential role in facilitating access to relevant datasets in the biosciences and environmental science.

This document presents the scope, approach, and findings of that study and recommendations for further action and research. The results are based on RAND's expertise and previous experience in this area and on preliminary investigation of a small set of potential candidate reference data collections.

The report is directly intended to inform the STM strategy at the British Library, but it should also be of interest to decision-makers at other national and research libraries faced with the challenges of the dynamic and complex STM landscape. It will also be relevant to other stakeholders in the research process, including researchers, funders, and organisations hosting, maintaining or governing data collections.

RAND Europe is an independent, private, not-for-profit research institution which helps to improve policy and decision-making through research and analysis. RAND Europe is an independently-chartered European unit of the globally operating think tank, the RAND Corporation.[1] For more information about RAND Europe or this document, please contact:

Stijn Hoorens
RAND Europe
Westbrook Centre
Milton Road
Cambridge CB4 1YG
United Kingdom
Tel: +44 1223 353329
email: hoorens@rand.org

Jeff Rothenberg
RAND
1776 Main Street
PO Box 2138
Santa Monica, CA 90407
USA
Tel: +1-310-393-0411
email: jeff@rand.org

[1] For more information about the RAND Corporation and RAND Europe, please see: http://www.rand.org and http://www.randeurope.org.

Contents

Table of Figures

Table of Tables

Abbreviations

API	Application Programming Interface
BL	British Library
CORBA	Common Object Request Broker Architecture
IP	Intellectual Property
J2EE	Java 2 Platform, Enterprise Edition
NSF	National Science Foundation
SOA	Service Oriented Architecture
STM	Scientific, Technical and Medical
ToR	Terms of Reference

Acknowledgments

The authors wish to acknowledge Neil Robinson (RAND Europe) and Neil Beagrie (Charles Beagrie Limited), for their critical and constructive comments on earlier versions of this document during the Quality Assurance process. Furthermore, we wish to thank Joachim Krapels and Lisa Cleverdon for their help during the publication process of this report. Finally, we are thankful to the British Library as the sponsor of this study, and particular to Lee-Ann Coleman and Elizabeth Newbold (British Library) for their helpful guidance.

Executive Summary

In recent decades, online access to large, high quality data collections has led to a new, deeper level of sharing and analysis, potentially accelerating and improving the quality of scientific research. These online datasets are becoming imperative at all stages of the research process, particularly in the areas of scientific, technical and medical (STM). Since libraries have a traditional responsibility to guarantee the availability of the output of scholarly research, they have a potentially important role to play in facilitating long-term access to these resources. Yet the role of a national library in the realm of STM data remains unclear.

This document presents the results of a scoping study that addresses the potential role of the British Library (BL) in facilitating access to relevant datasets in the biosciences and environmental science. The aim of this study is to assist the BL in developing an appropriate strategy that would enable it to establish a role for itself in the intake, curation, archiving and preservation of STM reference datasets, in order to provide access to these datasets for research purposes. The focus of this study is to explore a range of alternative strategies for the BL, which might be different for different types of databases or for data supporting different research fields or disciplines.

Characterising the dimensions of reference data collections

In order to develop a strategy aimed at providing access to these resources, a comprehensive picture should be developed of the inherent diversity in which research data are produced and offered. On the other hand, since the BL might function as a gateway to these resources, it is equally important to characterise the interests and needs of the potential users of these datasets. Therefore, we distinguished between the supply domain of datasets on the one hand, and the use of such data, ie, the demand domain, on the other. As illustrated in Table 1, a set of seven supply-side dimensions and a set of five demand-side dimensions have been developed. Each dimension has several attributes to allow for a characterisation of each candidate database and to delineate a set of options for the BL related to each attribute.

The identified attributes can have different values that represent the variation among the data collections' characteristics. We explored the online resources of a small sample of candidate data collections, and, to the extent possible, reviewed documentation about their ownership, management, data processing and validation methods, access mechanisms, query interfaces, browsing capabilities, metadata, etc. The identified dimensions, their

attributes and the span of plausible values on the supply and demand side are given in Table 1.

Table 1. Span of plausible attribute values in supply- and demand-side dimensions

	Dimension	Attribute	Attribute values
S1	Access	Restriction	none, role-based (eg, government, commercial, individual), location-or-affiliation-based (eg, by country, agency, professional society), by-registration, requiring unpaid-membership, paid-membership, use-payment (unlimited or by data-item, query, dataset, etc.)
		Access media	online-only, offline-only, on-or-offline
		Granularity	attribute, data-item, query-result, subset, dataset,
		Functionality	low, medium, high
		Software-requirements	generic, modifiable, free-download, server-resident-proprietary, proprietary
S2	Scale, dynamism, coverage and completeness	Scale	small, medium, large
		Dynamism of discipline	frozen, static, dynamic, volatile
		Dynamism (of database)	frozen, static, dynamic, volatile
		Temporal-depth	historical, current-only, multiple versions/editions, multi-
		Coverage	narrow, medium, broad
		Completeness	low, medium, high
		Collection-strategy	passive, active
		Processing	none, minimal, significant, intensive
		Validation	none, minimal, significant, intensive
		Timeliness	low, medium, high
S3	Disciplinary usage	Cross-discipline	no, somewhat, yes
		Disciplines	<discipline designations>
		Level of user support	low, medium, high
S4	Interface	User-interface	menu, graphical-selection, text-query, graphics-input
		Programmable-interfaces	no, server-support, framework-support, API
S5	Interoperability	Self-describing data	no, somewhat, yes
		Semantic transparency	no, somewhat, yes
		Linkage-to-other-collections	no, somewhat, yes
		Use of semantic standards	no, somewhat, yes
		cross-domain semantic cross-walks	no, somewhat, yes
		Programmable interfaces	non-existent, unique, standard, open
S6	Ownership, funding, governance, management and contributors	Reputation	low, medium, high
		Involvement	low, medium, high
		Accessibility	low, medium, high
		Funding-level	low, medium, high
		Funding-reliability	low, medium, high
		Governance-quality	low, medium, high
		Sustainability	short-term, medium-term, indefinite
S7	Attribution & IP	Attribution completeness	low, medium, high
		Attribution accuracy	low, medium, high
		Attribution granularity	low, medium, high
		Licensing, registration, agreements with owners	inapplicable, minimal, partial, complete
		End-user licensing	inapplicable, minimal, partial, complete
		Redaction/anomalisation of data	inapplicable, minimal, partial, complete

	Dimension	Attribute	Attribute values
D1	Research methodology, funding and stakeholder requirements	Required-access-granularity	attribute, data-item, query-result, dataset, database
		Required-metadata	low, medium, high
		Required-access-to-models	low, medium, high
		Methods	<method designations>
		Publication/distribution requirements	<various>

D2	Discovery methods	Search-engines Discovery metadata Other discovery resources	generic, specialised generic, specialised <indexes, catalogues, etc.>
D3	Query style	Expressivity Desired-interface Required-programmable-access	low, medium, high menu, graphical-selection, text-query, graphics-input low, medium, high
D4	Federation	Need-to-federate Required-metadata-support	low, medium, high low, medium, high
D5	Cross-disciplinary usage	Cross-disciplinary-usage Required-metadata-support	low, medium, high low, medium, high
D6	Timeliness and temporal access	Required-recency Required-timestamp-granularity Desired-update-method Required-temporal-access	low, medium, high low, medium, high asynchronous, time-stamped, transaction-based historical, current-only, versioned, multi-epoch, reconstructible

Bundles of strategic options for a national library

Analysis of the sample of candidate collections has led to the identification of a range of optional approaches that address each or a small set of salient attribute values. Examples of such options include: the BL should (or should not) hold a given dataset itself or should (or should not) develop and provide its own metadata and query or access mechanisms for a given dataset.

As an initial exercise for how the BL can develop a strategy for providing long-term access to these high quality reference data collections, we specified three exemplary clusters of attribute values, each of which characterises a class of databases. Each such attribute cluster defines a bundle of options that, taken together, can be considered a strategy.

1. The first cluster of attributes can be labelled as neutral: it represents the issues arising in the sample of databases that were investigated. For this cluster, the national library might consider providing transparent access to the data collections.
2. The second cluster of attributes represents a class of databases involving a complex, demanding set of requirements combined with relatively minimum support by the database itself. For this cluster, the national library might consider providing gateway access to the data collections.
3. The third cluster represents a class of databases involving a simple, undemanding set of requirements, combined with relatively good support by the database itself. These data collections have minimal access restriction, and their supporting mechanisms are relatively simple. For this cluster, the national library might consider providing transparent access to the data collections.

The three bundles of options associated with these attribute clusters should be considered indicative strategies rather than definitive ones. The 'demanding' and 'undemanding' clusters have been deliberately formulated as two extreme ends on a spectrum of plausible cases. The BL may choose different options, depending on its missions and policies. The set of options for the demanding and the undemanding clusters are illustrated in Table 2.

Lessons and next steps

The option bundles presented in this document are only a starting point. The BL's strategy with respect to any given database should be decided on the basis of an overall assessment of the importance and uniqueness of that database, its relevance to the BL's policies with

regard to STM data, and the BL's assessment of the degree to which users of the database would benefit from having the BL apply its own curatorial, preservation, or access resources to the database.

Although the limited resources of our study enabled us to obtain reasonable information for most supply-side attributes, details of ownership and funding (accessibility of owners, owner reputation, reliability of funding, etc) could in many cases only be inferred by our necessarily informal methods. Demand-side attributes were even harder to obtain; our values for most of these attributes are derived deductively rather than empirically. These need to be validated and revised based on future demand-side analysis.

The results of this study should therefore be replicated with greater depth and resources, using a larger number and wider range of sample databases augmented by demand-side input from researchers and user groups. The more in-depth examination should employ direct contact with database administrators, parent organisations, data processing managers, discipline-based organisations whose members use the database, and user communities. This should help fill in the supply-side attributes for each database as well as providing demand-side attributes, whose values were supplied largely by assumptions in the current study.

Table 2. Three cases of salient point clusters with associated option bundles

	Dimension	Attribute	1. Neutral case	2. Demanding case	3. Undemanding case
S1	Access	Restriction	Provide transparent access to the database		Provide transparent access to the database
		Access media	Provide online-only access to data	Transparently pass requests for offline data to original data publisher	Transparently pass requests for offline data to original data publisher
		Granularity	Ignore granularity as being an insignificant factor	Ignore granularity as being an insignificant factor	
		Functionality	Rely on the database's native access functionality	Provide reduced, simplified access functionality	Rely on the database's native access functionality
		Software-requirements	Rely on free software provided by the database	Obtain or recreate necessary software	Rely on free software provided by the database
S2	Scale, dynamism, coverage and completeness	Scale	Provide transparent access to minimise BL resource impact	Analyse resource impact of BL's re-hosting or acting as gateway	Assume any added resource requirements can be absorbed
		Dynamism of discipline	Provide transparent access to minimise BL resource impact	Interact with user groups to meet evolving functionality and volume demands	Assume relatively constant demand and functionality
		Dynamism (of database)	Provide transparent access to minimise BL resource impact	Provide transparent access to minimise BL resource impact	Assume low levels of data submission and no restructuring
		Temporal-depth	Rely on the database's own temporal storage and access	Rely on the database's own temporal storage and access	
		Coverage	Target broad coverage databases for their wide appeal		
		Completeness	Target complete databases for their wide appeal		Target complete databases for their wide appeal
		Collection-strategy			
		Processing			
		Validation	Rely on the database owner to certify the validity of its data	Perform additional validation	Rely on the database owner to certify the validity of its data
		Timeliness	Assume that delays will not be a significant factor for users	Monitor and evaluate timeliness when providing access	Assume that delays will not be a significant factor for users
S3	Disciplinary usage	Cross-discipline		Rely on native database cross-discipline support	
		Disciplines			
		Level of user support	Provide no additional user support for cross-disciplinary usage	Provide added user support for cross-disciplinary usage	Provide no additional user support for cross-disciplinary usage
S4	Interface	User-interface	Rely on native database interface	Provide a simplified interface to the database	Rely on native database interface
		Programmable-interfaces	Do not use programmable interface capabilities	Make limited use of programmable interface capabilities	Do not use programmable interface capabilities
S5	Inter-operability	Self-describing data			
		Semantic transparency			
		Linkage-to-other-collections		Virtually associate BL linkage information with the database	
		Use of semantic standards	Rely on the native semantics used by the database	Use semantic standards to re-encode and/or describe the database	Rely on the native semantics used by the database
		cross-domain semantic cross-	Rely on native semantic cross-walks (if any) provided by the database	Provide semantic cross-walks across relevant domains	Rely on native semantic cross-walks (if any) provided by the database

	Dimension	Attribute	1. Neutral case	2. Demanding case	3. Undemanding case
		walks			
		Programmable interfaces		Make limited use of programmable interoperability interface	
S6	Ownership, funding, governance, management and contributors	Reputation	Improve access to a high-quality database	Improve access to a high-quality database	Re-host an important database to improve its reputation
		Involvement			
		Accessibility	Minimise interaction with the owner while providing access	Minimise interaction with the owner while providing access	Rely on owner interaction to help provide access
		Funding-level			
		Funding-reliability	Assume the database will continue to be supported	Develop a backup strategy with and for the database and provide a "dark archive" for the database	Assume the database will continue to be supported
		Governance-quality		Minimise interaction with the governing body while re-hosting or serving as a gateway to the database	Interact with the governing body just enough to provide transparent access
		Sustainability	Assume the database will continue to function	Develop a backup strategy with and for the database	Assume the database will continue to function
S7	Attribution & IP	Attribution completeness		Target a database (partly) to make its IP data more complete	Target a database
		Attribution accuracy		Target a database at least in part to improve its IP data quality	Target a database at least in part because it has accurate IP data
		Attribution granularity			
		Licensing, registration, agreements with owners		Establish appropriate licensing, registration, and agreements	Rely on native licensing, registration, agreements with owners
		End-user licensing	Rely on native end-user licensing	Establish appropriate end-user licensing	Rely on native end-user licensing
		Redaction/anomalisation of data	Rely on native redaction/anomalisation	Provide added redaction and/or anomalisation facilities	Rely on native redaction/anomalisation

	Dimension	Attribute	1. Neutral case	2. Demanding case	3. Undemanding case
D1	Research methodology, funding and stakeholder requirements	Required-access-granularity	Ignore granularity as being an insignificant factor	Ignore granularity as being an insignificant factor	Ignore granularity as being an insignificant factor
		Required-metadata	Develop new metadata to support BL users	Develop new metadata to support BL users	Rely on existing metadata to support BL users
		Required-access-to-models	Rely on the database's native access to models	Rely on the database's native access to models	Rely on the database's native access to models
		Methods		Provide query mechanisms targeted to specific research methods	
		Publication/distribution requirements		Provide added support for publication and distribution	Rely on native database support of publication and distribution
D2	Discovery methods	Search-engines		Add support for specialised search-engines	
		Discovery metadata		Add specialised metadata to aid discovery	
		Other discovery resources		Create additional discovery resources for the database	
D3	Query style	Expressivity	Rely on the database's native query mechanisms		Rely on the database's native query mechanisms
		Desired-interface	Rely on the database's native interface	Interact with user groups to support their interface needs	Rely on the database's native interface
		Required-programmable-access	Do not provide access to programmable access capabilities	Provide access to any native programmable access capabilities	Do not provide access to programmable access capabilities
D4	Federation	Need-to-federate		Provide methods and advice for performing federation	
		Required-metadata-support	Rely on the database's native metadata to support federation	Rely on the database's native metadata to support federation	Rely on the database's native metadata to support federation
D5	Cross-disciplinary usage	Cross-disciplinary-usage		Rely on native database cross-discipline support	
		Required-metadata-support	Rely on the database's native metadata	Rely on the database's native metadata	Rely on the database's native metadata
D6	Timeliness and temporal access	Required-recency	Assume that recency will not be a significant factor for users	Monitor and evaluate recency when providing access	Assume that recency will not be a significant factor for users
		Required-timestamp-granularity	Rely on a database's native timestamp granularity	Avoid targeting a database due to inappropriate timestamp granularity	Rely on a database's native timestamp granularity
		Desired-update-method	Rely on a database's native update method		Rely on a database's native update method
		Required-temporal-access	Rely on the database's own temporal storage and access		Rely on the database's own temporal storage and access

CHAPTER 1 Background and introduction

1.1 A new paradigm for scientific research

The digital revolution has offered a myriad of new opportunities across nearly all sectors. The daily activities of librarians and archivists have benefited enormously from machine processing, database technology, increased storage capacity, and electronic content delivery, to name just a few obvious examples.

In the STM fields, there has been a trend toward the increasing use of data and models underlying published results. Scientists are disseminating and using each other's datasets, models, lab notes, etc, to help analyse, reproduce, and review each other's work (reproducibility being essential to the scientific method). This is seen by many as a significant paradigm shift to a new, deeper level of sharing and analysis, which has the potential to greatly accelerate and improve the quality of scientific research.

Moreover, in many cases (such as the Human Genome Project and space-based earth observation projects), the primary results of a scientific project are data rather than analytic reports. In such cases, although published reports may summarise the results, the valued output of the project lies in the data that it produces. Such products may take various forms, but they are most likely to consist of databases, including pre-packaged queries and executable programs that provide appropriate and useful access to the data.

These developments have been endorsed by various international organisations and decision makers. In January 2004, the Organisation for Economic Co-operation and Development (OECD 2004) proposed ten principles for open access to research data from public funding. The European Commission recently discussed the future of scientific publishing as well, and committed to financially support the establishment of digital repositories for storing scientific data and digital preservation initiatives (European Commission 2007). Finally, in the UK, the Research Councils (RCUK 2005) have published a statement presenting their position with regard to (open) access to research outputs.

1.2 Access to research data reference collections

The digital revolution has given way to a myriad of digital data collections, involving many participants using many types of data for many different purposes. This report focuses on

those digital data collections that merit the facilitation of long-term access for the STM research community. The data collections that are potentially of most value in this regard serve large segments of the STM research community and conform to robust, well-established and comprehensive standards. In a recent publication by the U.S. National Science Foundation (NSF 2007) these collections are referred to as reference collections, using a taxonomy introduced by the National Science Board (NSB 2005). As specified in the Terms of Reference for this study, these collections form the prime focus of this study. A second class of data collections is referred to as resource or community collections. These collections usually serve a single science or engineering community, and they can be of value for the wider STM research community. Access to a third class of data collections, those created by individual investigators and investigator teams and serving an immediate group of participants, may have lower priority. The NSF (2007) classifies such collections as research collections. It is important to recognise that the digital universe of data collections is not only heterogeneous and complex, it is also dynamic. Collections may grow, evolve and migrate from one class to another over time.

Data centres may also consider supporting long-term access to small supporting data sets that are published concurrently with, or are an electronic component of, journal articles. These supporting data collections are not explicitly distinguished by the three classes above. Although supplementary datasets are not a priori excluded, we primarily focus on collections created by a community of researchers, rather than a small set of co-authors or project members.

As noted by the NSF (2007), the research community has witnessed the rise of a multitude of collections that are robust and flexible, while allowing for heterogeneous data types and associated metadata, enabling them to meet the wide range of needs, practices and expectations that are found among the communities of data authors and users. Data centres must accommodate this heterogeneity in order to support long-term access to these data collections and enable effective research in a digital environment.

1.3 The functions and services of STM data collections and centres

The increasing use and relevance of electronic data collections is now in widespread use throughout most research disciplines and have opened up new avenues for collaboration and sharing. A service that identifies data sources, provides access and enables further use and manipulation of the data, will become an increasingly important resource for researchers.

However, there are important differences between and within disciplines with regard to funding, data access policies and practices Not only do some disciplines rely more heavily on access to electronic data collections than others – for example biomedical informatics versus theoretical mathematics – it is also not difficult to imagine that the differences between large scale nuclear physics experiments with a particle accelerator and genome sequencing the complete gene profile of an organism have important consequences for the way in which these data are stored, labelled, archived, managed, accessed, manipulated, etc. As a consequence of these disciplinary differences, the benefits from improved access to

research data will also deviate per research community, and so will the added value of data centres.

Major existing STM data centres can perform a number of functions. Many serve as collection "funnels" for data in their respective scientific fields: researchers submit data to these repositories because they know that doing so will make their results visible to and accessible by fellow researchers. Such services also perform basic categorisation and structuring of their contents, along with the creation of various kinds of indexing, search tags, and other finding aids. Finally, they may provide access to their data by means of web-based query interfaces and browsing tools, which may utilise graphical input, ontologies, keywords, and many other techniques to allow users to find what they need.

In a study investigating the relationships between data centres and institutions which may develop data repositories (Lyon 2007), the University of Bath (UKOLN) outlined the possible functions of data centres:

- Managing data for the long-term;
- Meeting standards for good practice;
- Providing training for deposit;
- Promoting the repository service;
- Protecting rights of data contributors; and
- Providing tools for re-use of data.

Data Collections in some disciplines including crystallography, space and earth observation, or social sciences for example have been in existence now and maintained for several decades. They have well-defined community standards and subject data centres but other areas; for example, many areas of biology, are more fluid. There are different retention and use periods for different datasets so not all need to be preserved in perpetuity. There are different preservation requirements for collections of closed datasets and dynamic datasets that are still being continuously updated. Collections can be transferred between institutions over time so sustainability of a data centre (ie, a service provider for a set of functions for a data collection) can be separated out to some degree from the preservation/sustainability of the data collection.

As guardians of the scholarly record and facilitators of access to academic resources, libraries have a unique position in this field. Given the substantial differences between and within disciplines, it may become increasingly difficult to capture the most important collections in a consistent manner with a view of promoting interdisciplinary research. Hence, there could be a role for organisations, such as national libraries, to act as a central aggregator. However, the UKOLN study did not discuss the implications for the role of libraries in facilitating these data centres.

1.4 The role of a national library is so far unclear

As custodians of the collective memory, libraries have taken the role of preserving and facilitating access to scholarly research output. As the availability of research data has

become an imperative aspect of all stages in the research process, national libraries have a potentially important role to play in facilitating such access to these resources. Yet the role of a national library in the realm of STM data is so far unclear.

Ingestion, collection, curation, validation, and the provision of access to such data goes well beyond the traditional role of a library. Yet the ever-increasing value of STM databases makes them a crucial scientific resource, the perennial accessibility of which may be too important to leave to the sole responsibility of their owners. It may therefore be appropriate for national libraries such as the BL to define a role for them in ensuring the continued preservation and accessibility of STM data. Such a role might take various forms. At one extreme, a national library might simply provide an alternative access channel to data stored in existing databases, by making each database's own access mechanisms and interfaces visible to the library's users. At the other extreme, the library might replicate and re-host a database, creating its own copy of the data, managing and preserving this copy itself, and providing access to it by mechanisms that may or may not be similar to those provided by the original database. In between these extremes lie a number of intermediate possibilities. These roles are not solely dependent on the technical specifications of the data collections and the technical user requirements, however. There is a wide stakeholder community with varying interests and views on the functions of data centres and the role of a national library therein.[2]

The national library might provide backup for data in case a database becomes inaccessible or cannot be sustained. Or it might serve as a gateway to an existing database. This would add value to the database by providing its own data quality enhancements, metadata, search facilities, access mechanisms, federation tools for combining multiple databases to perform research, or other capabilities. Finally, data centres may also consider supporting long-term access to small supporting data sets that are published concurrently with, or are an electronic component of, journal articles. As libraries have a traditional role in the archiving and preservation of scholarly publications, this would be a natural expansion of the national library's remit. Initiatives that illustrate how such functions can be implemented include:

- The MRC Mental Health Cohort Directory.[3] This directory, funded by the MRC and the UK Mental Health Research Network facilitates the discovery of existing cohort study datasets of relevance to mental health research. It provides a high-level description of each dataset's characteristics including a statement on data access and a link to the originating study websites. Its initial content is based on an MRC commissioned survey.

[2] For possible stakeholder concerns see for example: Bryan Lawrence... personal wiki, blog and notes. "Why not dinosaurs?" by Bryan Lawrence: 2007/10/12. Available at: http://home.badc.rl.ac.uk/lawrence/blog/2007/10#fn3call

[3] See: MRC and UK Mental Health Research Network. "MRC and MHRN Data Discovery Resource: Population Studies of Relevance to Mental Health Research". Accessed on: 3/3/2008. Available at: http://www.mrc.ac.uk/consumption/groups/public/documents/content/mrc003776.pdf

- The National Virtual Observatory.[4] The NVO is an initiative sponsored by the National Science Foundation with the Johns Hopkins University, developed in collaboration with the International Virtual Observatory Alliance. It consists of a set of online tools to link worldwide astronomy data together, enabling access to data from different instruments. The NVO site is a gateway to education and public outreach resources from the NVO intended for students, teachers, and the public.
- Intute.[5] Intute is a free online service providing access to online resources for education and research. Intute selects material through a network of subject specialists to create a database. This aggregates a large range of resources and facilitates access to both subject-specific and cross-subject resources, all of which have been evaluated for their quality and relevance.

In this document we present the results of a scoping study that addresses the potential role of the BL in facilitating access to relevant datasets in the biosciences and environmental science. The aim of this study is to assist the BL in developing an appropriate strategy (or a set of alternative tactics) that would enable it to establish a role for itself in the intake, curation, archiving and preservation of selected STM reference datasets, in order to provide access to and manipulation of these datasets for research purposes. The focus of this study is on this range of alternative strategies for the BL, which might be different for different types of databases or for data supporting different research fields or disciplines.

1.5 Structure of this report

This report is structured in seven chapters. The next chapter, Chapter 2, outlines the scope of this study and explains the five steps that we have followed towards developing recommendations to enable the BL to establish a role for itself in providing access to and manipulating STM reference data collections. These five steps are represented by the five remaining chapters. Chapter 3 delineates the aspects of the supply-side of reference data collections that are relevant when considering a role for the national library, while Chapter 4 addresses the corresponding aspects for the demand-side. In Chapter 5, we characterise a sample set of four data collections and identify the range of variation across these collections on the different supply- and demand-side dimensions. Chapter 6 synthesises the findings from previous chapters and the consequences for the BL, identifying a range of appropriate bundles of options. Based on these option bundles, we developed a set of three alternative strategies that the BL could take. These strategies are also outlined and explained in Chapter 6. Chapter 7, which concludes the report, delineates lessons from this approach, and identifies issues that need to be accounted for if a national library is to define a role for itself in archiving, preserving and providing access to STM reference data collections for research purposes.

[4] See: John Hopkins University. "National Virtual Observatory". Accessed on: 3/3/2008. Available at: http://www.virtualobservatory.org/

[5] See: Burgess (2006).

CHAPTER 2 Scope and approach

2.1 Defining some terms

Various terms are recurring frequently in this report. Some terms, such as data collection and dataset, can be used interchangeably. Although closely related, there are subtle but important differences between most of the other concepts. In this section, we briefly define some of the most important terms used in this report.

Data. For the purposes of this document, data can refer to any information that can be stored in digital form, including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc. These data entities may be produced by observations, experiments, simulations and models, along with the associated documentation needed to describe and interpret those data (NSF 2007).

Dataset. We use the term dataset to mean a specific set of data, as produced by a single study, experiment, event, etc., rather than a database, which may contain many such datasets in a related field.

Database. We use the term database to mean a dataset or a collection of multiple datasets in a digital format that can be accessed through a common access and/or management system.

Data collection. We use the term data collection to mean a collection of datasets that can be accessed through either a common or through separate access and/or management system. The terms data collection and database overlap when they refer to a collection of datasets that have a common access and management system. Therefore, the two concepts are often used interchangeably. The NSF uses the term 'collection' to refer not only to stored data but also to the infrastructure, organisations, and individuals necessary to preserve access to the data.

Gateway access. We use the term gateway access to mean providing access to a data collection via some alternative site (such as the BL) other than the collection's home site, without reproducing or re-hosting the collection or its fundamental access mechanisms. A gateway may, however, provide additional access or query mechanisms or additional curatorial or explanatory services beyond those offered by the collection's home site.

Metadata. Metadata are a subset of data, and are data about data. Metadata summarise data content, context, structure, interrelationships, and provenance (information on

history and origins). They add relevance and purpose to data, and enable the identification of similar data in different data collections (NSF 2007).

Research Collections. Authors are individual investigators and investigator teams. Research collections are usually maintained to serve an immediate group of participants only for the life of a project, and are typically subjected to limited processing or curation. Data may not conform to any data standards (NSF 2007).

Resource or Community Collections. Resource collections are authored by a community of investigators, often within a domain of science or engineering, and are often developed with community level standards. Budgets are often intermediate in size. Lifetime is between the mid- and long-term (NSF 2007).

Reference Collections. Reference collections are authored by and serve large segments of the science and engineering community and conform to robust, well-established and comprehensive standards, which often lead to a universal standard. Budgets are large and are often derived from diverse sources with a view to indefinite support (NSF 2007).

The latter three concepts are not fixed over time. Data collections can be dynamic and may grow when it is populated with more data and complexity is added. As a consequence, research collections that are initially primarily maintained to serve an immediate group of participants, may become used by a whole community, and eventually serve large segments of the STM fields. An audit and classification of existing data collection can therefore only be interpreted as a snapshot of a certain moment in time.

Transparent access. We use the term transparent access to mean providing alternative access to a data collection without reproducing or re-hosting the collection or its access mechanisms, and without providing any additional access or query mechanisms or additional curatorial or explanatory services beyond those offered by the collection's home site.

2.2 A five-step approach

In our approach we have distinguished five distinct steps towards developing recommendations to enable the BL to establish a role for itself in providing access to and manipulation of STM reference data collections for research purposes. These five steps are outlined in Figure 1.

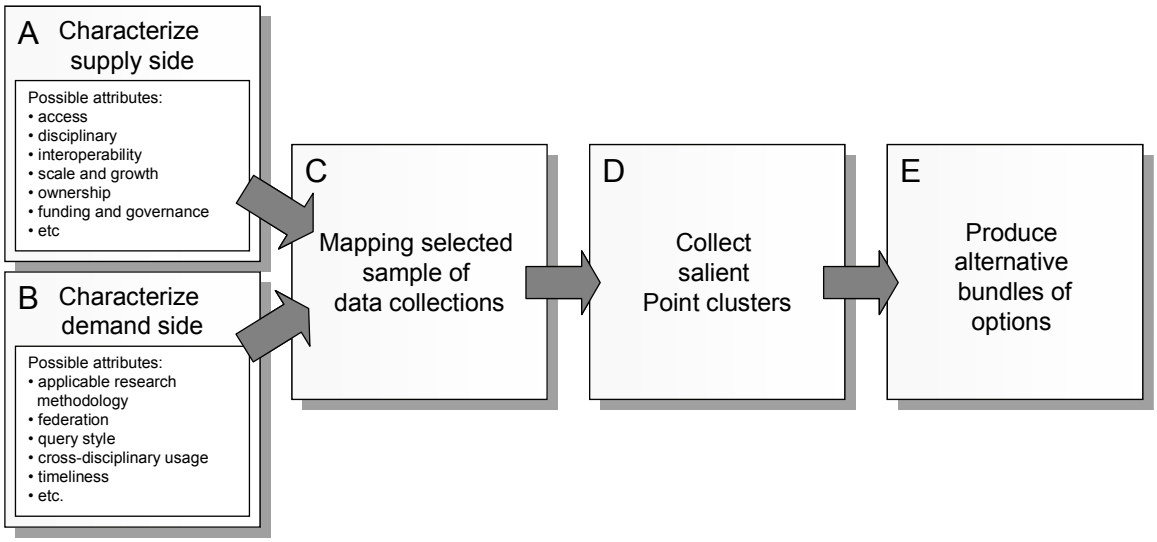

Figure 1. Schematic representation of proposed analytic framework.

2.2.1 Step A and B. Supply side and demand side typology

The basic framework for this analysis involved the development of a set of attributes that characterises each database and a set of options for the BL related to each such attribute. For this study it is important to make a distinction between the supply domain of datasets on the one hand, and the demand domain, i.e. the use of such data, on the other. While it is impossible to isolate the characteristics of the supply-side of data collections from those of the demand-side for these collections, distinguishing these sides is useful. In order to develop a strategy aimed at providing access to these resources, a comprehensive picture should be developed of the inherent diversity in which research data are produced and accessed. On the other hand, since the BL might function as a gateway to these resources, it is equally important to characterise the interests and needs of the potential users of these datasets.

2.2.2 Selection of a small sample of reference data collections

In order to delineate the potential variety of characteristics of data collections and the preferences of users using these collections, we have investigated a set of candidate collections. The sample data collections that were examined for this study were selected by a multi-step filtering process. We used a report prepared for the British Library by Key Perspectives (KP, 2007) as an initial source. This report characterises 227 databases in terms of their amount and quality of metadata and their overall quality.

The BL indicated that it is not interested in all data collections identified by the KP report (see Table 3). For example, publication tools, bibliographic databases or teaching materials are not of direct importance to the BL's STM strategy. Although many of the databases in the KP report contain a mixture of bibliographic and scientific data, the latter was the primary focus of the BL's interest for this study. As a consequence, we only considered those collections characterised by KP as: Biological pathway, specific bio-molecules (B); Specific discipline (e.g. toxicology) or system (e.g. embryology) or higher-order activity (D); Genetics or genetic system (G); Mathematical models, modelling techniques,

analytical tools, data derivation tools, experimental types (M); Single gene (O); Taxonomy, seeds, specimens, phylogenetic data (T); or Proteins or proteomics (X).

Because the definitions of "overall score" and "metadata score" in the KP report overlap and are not independent of each other, they do not constitute independent evaluation criteria. Nevertheless, the combination of the two criteria provides a credible measure of the overall quality of each database. The next filtering step therefore consisted of selecting those databases with the highest quality scores in that report. This produced a set of relatively high quality candidate databases. It must be noted that this approach may have several inherent selection biases. For example, the resulting sample of data collections is likely to be biased towards already well-funded and curated databases, for which a data centre may have less added-value. This needs to be taken into account when interpreting the results of this preliminary investigation.

Table 3. Data collection typology used as basis for selection

Code	Database content type (predominant one if mixed)	Of interest to BL
A	Teaching materials	
B	Biological pathway, specific biomolecules	√
D	Specific discipline (e.g. toxicology) or system (e.g. embryology) or higher-order activity	√
G	Genetics or genetic system	√
I	Images	
J	Publication tools	
L	Literature or bibliographic	
M	Mathematical models, modelling techniques, analytical tools, data derivation tools, experimental types	√
O	Single gene	√
P	Portal or cross-database search service	
T	Taxonomy, seeds, specimens, phylogenetic data	√
X	Proteins or proteomics	√

SOURCE: Key Perspectives (2007)

We then performed a cursory examination of each of the databases in the resulting set, examining their websites in order to narrow the sample to a relatively small number of databases in each category (biological sciences, environmental sciences, physics, and chemistry) that spanned a range of different data types, user interfaces, and ownership characteristics. We further reduced this based on the amount of documentation for each database that appeared to be available on its website. This produced a set of 13 "preferred" candidates.

Finally, we narrowed the resulting set of candidates to as small a number as possible, while retaining a range of categories, data types, user interfaces, and ownership characteristics. This yielded the final set of 4 target databases that we studied:

1. Allen Brain Atlas
2. PubChem
3. FishBase
4. PDBsum

Dimensions

We have delineated a set of supply and demand side dimensions that are relevant when considering enabling access to these reference collections. These dimensions are based on initial investigation of the above candidate reference collections, as well as RAND's background in scientific database development, management and usage. An initial set of dimensions was then discussed with the BL. In accordance with its feedback, closely related dimensions were clustered and missing dimensions were added subsequently. The set of supply- and demand-side dimensions are listed in Box A and Box B and are discussed in more detail in Chapter 3 and Chapter 4, respectively.

Box A. Supply-side dimensions

- Access
- Scale, dynamism, coverage and completeness
- Disciplines
- Interface
- Interoperability
- Ownership, funding, governance, management and contributors
- Attribution & IP

Box B. Demand-side dimensions

- Research methodology, funding and stakeholder requirements
- Discovery methods
- Federation
- Query style
- Cross-disciplinary usage
- Timeliness and temporal access

Attributes

Each dimension has a set of attributes. For example, there are different aspects of the dimension "Access" that are worth considering for a potential data centre. These include the need for authorisation, membership, or payment before accessing data, restrictions on who can access data, the need for specialised software to access data, etc. Attributes were derived from RAND's previous experience and our analysis of the small sample of data collections. As with the dimensions, the initial set of attributes was then discussed with the BL and subsequently adjusted.

As we studied the selected data collections in our small sample, we attempted to characterise them in terms of the above supply- and demand-side attributes. In addition, we adjusted or modified the attributes themselves as we learned more about the databases. The resulting characterisations for each supply- and demand-side dimension are outlined in Chapter 3 and Chapter 4, respectively. These are somewhat tentative, due to the fact that our exploration of each database was limited to what we could learn from its website.

Furthermore, this characterisation is commendable in its focus on information management and focuses on characteristics and structures associated with information held in existing data collections. The technical aspects of their functionality (eg, Relational Database Management Systems, 3rd generation XML based datastores, etc.) are not covered in this characterisation. Nevertheless, we believe that our framework provides a reasonable first approximation to a useful description of each database for the purpose at hand.

Attribute values

The identified dimension attributes can have different values, which correspond to the variation among the data collections' characteristics (supply) and the different preferences of researchers using these data collections (demand). We explored the online resources of each data collection, and, to the extent possible, reviewed documentation about its:

- ownership;
- management;
- data gathering or generating strategies and procedures;
- data processing and validation methods;
- access mechanisms;
- query interfaces;
- browsing capabilities;
- dataset download capabilities;
- metadata;
- linkages to other databases;
- user-support; and
- server-provided software, downloadable software, and application programming interfaces (APIs).

We also followed links providing information about parent organisations, funders, scientific review boards, user communities, publications describing the database, and online user forums. Finally, we performed web searches to try to find usage information about each database and to glean some sense of its reputation in the relevant scientific disciplines. This enabled us to specify a range of plausible values for each attribute, including some intermediate values.

2.2.3 Step C. Mapping the sample of selected databases

In this step we mapped the four sample databases onto dimensions and their attributes identified in the previous steps. We investigated websites of these databases and retrieved their attribute values through following the links to background information, related sites, ownership and funding organisations. The results of this mapping procedure are illustrated in summary tables. Finally, we overlaid the four collections in our sample to get an indication of the potential span of attribute values of all candidate reference data collections.

2.2.4 **Step D. Salient point clusters**

Initially, our aim was to develop "salient point-pairs" of supply- and demand-side attributes that would correspond to important and relevant cases for the BL. We would identify a small but representative set of examples of points along each dimension, as well as what appeared to be the most salient pairs of such points from the supply and demand (dataset and usage) domains. However, as our study proceeded, we realised that individual point-pairs of this kind provide a weaker characterisation of a database than more robust clusters of multiple attribute values. We therefore modified our approach to instead create "point-clusters" consisting of values for multiple supply and demand attributes. Not every attribute in each such cluster has a value, but the values present in each cluster can be thought of as providing a "signature" for a class of databases. We developed three such clusters, the first of which corresponds to a "neutral" case based on the four sample databases we examined, whereas the second and third clusters represent extremes, which we labelled "demanding" and "undemanding" cases.

Each attribute value was also mapped to a proposed option (or set of alternative options) that would enable the BL to address the combination of supply and demand attributes represented by that point-cluster. Examples of such options might be that the BL should (or should not) hold a given dataset itself or should (or should not) develop and provide its own metadata and query or access mechanisms for a given dataset.

2.2.5 **Step E. Develop alternative strategies**

Each of the relevant attributes in each cluster is mapped to one of the BL options for that attribute. The cluster as a whole therefore specifies a set of options which, taken together, can be considered a strategy. A given attribute value does not uniquely determine a single option: instead, the entire set of values for all attributes in a cluster provides the context in which a given attribute value is mapped to a given option. Even the entire context of the cluster does not uniquely determine an option: the BL may choose different options in the same context depending on its missions and policies.

For example, a database that is considered important but whose quality or support is not very good might lead to two quite different strategies--one in which the BL chooses not to associate itself with the database to avoid tarnishing the BL's reputation, and the other in which the BL chooses to take on the task of improving the quality of the database and providing access to it under the BL's own aegis.

The three bundles of options presented in this report should be considered indicative rather than definitive. We feel strongly that the results of this study should be replicated with greater depth and resources, using a larger number and wider range of sample databases augmented by demand-side input from researchers and user groups and by additional input from the BL itself on its perceived options and overall strategies for dealing with scientific data.

Finally, we have not assessed the feasibility or implementability of these options; for example, cost considerations have not been within the remit of this preliminary study. The BL options for each attribute should be validated, refined and feasibility should be further assessed by BL staff. Therefore, the options can be expected to evolve as the BL continues

its exploration of the issues associated with re-hosting or providing alternative access to scientific databases.

In addition to the results of this preliminary investigation, we have formulated a series of lessons drawn from applying this methodology. Based on these, we discuss the resources that would be required for the BL to apply this characterisation to all potential candidate reference data collections. Finally, we outline the steps that are necessary to undertake a more comprehensive assessment of researchers' preferences.

CHAPTER 3 Supply-side characterisation

In this chapter we discuss a set of supply-side dimensions that are relevant when considering enabling access to these reference collections. We identify the attributes of these dimensions and outline plausible values of these attributes based on an investigation of a small sample of data collections.

1. Access
2. Scale, dynamism, coverage and completeness
3. Disciplines
4. Interface
5. Interoperability
6. Ownership, funding, governance, management and contributors
7. Attribution & IP

This set of dimensions and the specific attributes that we identify within each dimension are the most obvious ones that emerged from our sample set of data collections. However, they are by no means exhaustive, and further analysis of these and other data collections will no doubt reveal additional attributes and possibly additional dimensions that may also be relevant.

3.1 Access

Access attributes include the need for authorisation, membership, or payment before accessing data, restrictions on who can access data, the need for specialised software to access data, etc. In addition, these include the availability of online access and/or offline access, eg, via the distribution of physical storage media, such as tapes or DVDs. Access attributes also include the granularity of the information that is available in a collection. The range of plausible attribute values includes:

Restriction:	none, role-based (e.g., government, commercial, individual), location-or-affiliation-based (e.g., by country, agency, professional society), by-registration, requiring unpaid-membership, paid-membership, use-payment (unlimited or by data-item, query, dataset, etc.)
Access media:	online-only, offline-only, on-or-offline

Granularity:	attribute, data-item, query-result, subset, dataset, database
Functionality:	low, medium, high
Software-requirements:	generic, server-resident, modifiable, freely downloadable, proprietary

3.2 Scale, dynamism, coverage and completeness

The scale and growth pattern of a database affect its stability, its continuing relevance, the size of its potential user community (and therefore indirectly its support base), etc. Dynamism can characterise the underlying discipline or disciplines that the database represents (dynamism of discipline) as well as the database itself (dynamism of database). Both of these characterisations, as well as the relationship between the two, are relevant to users: a frozen or static database representing a dynamic or volatile discipline may not provide up-to-date data.

The temporal depth of a database concerns the degree to which it provides historical data, past versions of its data or data from multiple epochs versus current data.

The coverage of a collection may be only implicit in its data. Users may need to know what is not contained in a database (ie, what is excluded) as well as what is contained in it. Coverage may be implicit in a discipline or sub-discipline, but it may also consist of the explicit scope of data within a discipline or the scope of individual datasets within a database. Coverage is also relative to the dynamism of the underlying discipline, whose scope may be expanding over time. Similarly, completeness is relative to the dynamism and completeness of the discipline, as well as the temporal depth of the database. A dynamic database representing a volatile discipline may at best be complete as of the current time.

The collection strategy of a database may be passive – serving as a repository allowing practitioners in a discipline to insert data – or active – aggressively seeking data for the database. Processing and validation may vary, affecting the value and quality of the database. Finally, timeliness is a measure of how up-to-date the database is, regardless of its temporal depth or completeness.

The attributes for this dimension and their associated range of plausible values can be listed as follows:

Scale:	small, medium, large
Dynamism of discipline:	frozen, static, dynamic, volatile
Dynamism (of database):	frozen, static, dynamic, volatile
Temporal-depth:	historical, current-only, multiple versions/editions, multi-epoch
Coverage:	narrow, medium, broad
Completeness:	low, medium, high
Collection-strategy:	passive, active
Processing:	none, minimal, significant, intensive

Validation:	none, minimal, significant, intensive
Timeliness:	low, medium, high

3.3 Disciplines

These attributes specify the discipline or disciplines that are represented by the data in a collection and the degree to which the collection is cross-disciplinary. This in turn may imply assumptions about the meaning of data, the methods and basis used in collecting, measuring, computing, and processing data, the interpretation of data, the kinds of metadata provided, the level of user-support required and provided, etc. The range of plausible attribute values includes:

Cross-disciplinary:	no, somewhat, yes
Disciplines:	<discipline designations>
Level of user support:	low, medium, high

3.4 Interface

These attributes describe the user interface that a database provides to its users and the degree to which the database makes its functionality available via an application programming interface (API), as part of a distributed computing framework (such as J2EE or CORBA), as a Web Service, in a Service Oriented Architecture (SOA) paradigm, etc. The range of plausible attribute values includes:

User-interface:	text-menu, graphics-menu, text-input, graphics-input
Programmable-interfaces:	no, server-support, framework-support, API

3.5 Interoperability

These attributes involve both technical and semantic interoperability. Technical issues include the format of the data, the degree to which data fields are self-describing, and the ability of the database to perform automated conversion into other formats, other units, or other bases. Technical attributes also concern linkages from a given collection to other collections, as well as user interface and software issues, such as the degree to which the database's programmable interfaces (API, distributed computing framework, Web Service or Service Oriented Architecture (SOA) paradigm, etc.) are standardised and interoperable with those of other databases and of relevant query and analysis software. Semantic interoperability issues include the compatibility of the underlying models used to generate, process, and analyse data for different purposes, as well as the use of documentation standards, cross-walks between different domain standards etc.

The range of plausible attribute values includes:

Self-describing data:	no, somewhat, yes
Semantic transparency:	no, somewhat, yes

Linkage-to-other-collections:	no, somewhat, yes
Use of semantic standards:	no, somewhat, yes
Cross-domain semantic cross-walks: no, somewhat, yes	
Programmable interfaces:	non-existent, unique, standard, open

3.6 Ownership, funding, governance, management and contributors

These attributes affect a collection's credibility, reliability, level of support, and the kinds of capabilities and interfaces it is likely to offer to various user communities. Governance also affects how responsive a collection is likely to be to evolving user needs. In some cases, it is important to distinguish ownership, funding, governance and management from each other, as well as to identify the community of data contributors. For each of these roles, the range of plausible attribute values includes:

Reputation:	low, medium, high
Involvement:	low, medium, high
Accessibility:	low, medium, high
Funding-level:	low, medium, high
Funding-reliability:	low, medium, high
Governance-quality:	low, medium, high
Sustainability:	short-term, medium-term, indefinite

3.7 Attribution & IP

This concerns the ways in which a collection attributes its data to proper sources and tracks intellectual property (IP) rights, both for legal purposes and to ensure proper pedigree for the data in question. Ownership of the database itself may be distinct from ownership of IP rights to its contents. These attributes represent the degree to which the ownership, privacy and confidentiality of data are legally, ethically, and procedurally protected. The range of plausible attribute values includes:

Attribution completeness:	low, medium, high
Attribution accuracy:	low, medium, high
Attribution granularity:	low, medium, high
Licensing, registration, agreements with owners: inapplicable, minimal, partial, complete	
End-user licensing:	inapplicable, minimal, partial, complete
Redaction/anomalisation of data: inapplicable, unavailable, as-needed	

CHAPTER 4 Demand-side characterisation

There are various dimensions defined by the users of datasets that influence the way in which these datasets are used. In this chapter, we discuss an initial set of relevant demand side dimensions:

1. Research methodology
2. Discovery methods
3. Query style
4. Federation
5. Cross-disciplinary usage
6. Timeliness and temporal access

The distribution of reference collections across these dimensions determines how the British Library could facilitate access to these datasets. These dimensions are discussed in more detail below. For each demand-side dimension we have outlined several relevant attributes, and we have specified the range of plausible attribute values, including some intermediate values.

Because this study did not have sufficient resources to explore demand-side issues in any depth, this set of dimensions and the specific attributes that we identify within each dimension are those that intuitively appeared relevant to our sample set of data collections. However, further analysis of demand-side issues and user concerns surrounding these and other data collections will no doubt reveal additional attributes and possibly additional dimensions that may be equally relevant.

4.1 Research methodology, funding and stakeholder requirements

The approaches that users of data employ to do their research have implications for the ways they may prefer databases to be organised and accessed. For example, survey methods may need to compare multiple datasets from multiple studies, whereas methods that focus on specific subsets of data may need to analyse individual datasets in depth.

Research methods may affect the kinds of metadata that users need, the kinds of indexing, search, and mining tools they need, and the granularity of access they require. In addition, some methods may require access to underlying models used to generate or process data. Methods may be different for users in different disciplines. Finally, researchers may need to

satisfy requirements or other constraints levied on them by the funders of their work or other stakeholders, such as the need to share or distribute their results within specific communities, deposit them in specific repositories, or publish them in specific venues. The range of plausible attribute values includes:

Required-access-granularity:	attribute, data-item, query-result, dataset, database
Required-metadata:	low, medium, high
Required-access-to-models:	low, medium, high
Methods:	<method designations>
Publication/distribution requirements:	<various>

4.2 Discovery methods

This is related to research methodology, but has to do with how researchers search for and discover data resources. This may involve the use of generic or special-purpose search engines, metadata or meta-databases, citation indexes, specialist catalogues, etc. The range of plausible attribute values includes:

Search-engines:	generic, specialised
Discovery metadata:	generic, specialised
Other discovery resources:	<indexes, catalogues, etc.>

4.3 Query style

This is related to research methodology, but it is a lower-level issue, having more to do with the specific mechanisms that users will employ to find and access data. This impacts the query-language design of a database, as well as its metadata, which should support various types of queries. For example, "semantic queries" may need to be able to ask about contexts, which may need to be supplied by metadata. The range of plausible attribute values includes:

Expressivity:	low, medium, high
Desired-interface:	text-menu, graphics-menu, text-input, graphics-input
Required-programmable-access:	low, medium, high

4.4 Federation

Users in some disciplines (and doing some kinds of research) may need to access data across multiple databases, thereby requiring federated schemas and federated metadata.[6]

[6] Examples of initiatives that provide access to multiple databases and thereby require such federation include: MRC Mental Health Cohort (see: http://www.mrc.ac.uk/consumption/groups/public/documents/content/mrc003776.pdf) and National Cancer Informatics Initiative (see: http://www.cancerinformatics.org.uk/).

Individual databases that are frequently used as part of a larger federation may be of limited value to users without access to other databases in the federation. This is largely a demand-side issue, though it could be a supply-side dimension as well. Federation is something that may fall outside the purview of any single database, so it may require some overarching responsibility, such as might be provided by the BL or by professional or disciplinary organisations. The range of plausible attribute values includes:

Need-to-federate:	low, medium, high
Required-metadata-support:	low, medium, high

4.5 Cross-disciplinary usage

This is related to federation, but it entails additional issues of cross-disciplinary translation. For example, basic definitions, units of measure, measurement assumptions, calibration techniques, etc, may differ across disciplines, so access of multi-disciplinary data may require additional mechanisms to align data for a given research purpose. It may be necessary to understand specific cross-disciplinary usage methods and patterns in order to account for its impact on database design and access requirements. The range of plausible attribute values includes:

Cross-disciplinary-usage:	low, medium, high
Required-metadata-support:	low, medium, high

4.6 Timeliness and temporal access

This is a function of the type of research that is being done and of the needs of each discipline. For example, public health may need rapid access to dynamic data to cope with epidemics, whereas astronomers may need to be alerted to ephemeral events. Timeliness is a function of the update rate of a database but also affects the need for metadata to provide timestamps and relevance or expiration dates on data. The range of plausible attribute values includes:

Required-recency:	low, medium, high
Required-timestamp-granularity:	low, medium, high
Desired-update-method:	asynchronous, time stamped, transaction-based
Required-temporal access:	historical, current-only, versioned, multi-epoch

CHAPTER 5 Mapping sample of data collections

This chapter presents the mapping of the attribute values of the four sample databases. Section 5.1 until 5.4 illustrate the characterisations of the four individual databases and present the attribute values of these data collections. In Section 5.5 we present the combined values of these four databases. As such, it is an attempt to define a class of databases that encompasses these and similar cases.

5.1 Allen Brain Atlas

The Allen Brain Atlas is a project of the Allen Institute for Brain Science in Seattle. It constitutes an 'atlas' of all known genes in the rat brain including their locations and other functional and molecular biological data.

Managed by: Allen Institute for Brain Science, Seattle

Covers: 85 million records of gene data (21,000 genes) and brain scans

URL: http://brain-map.org/

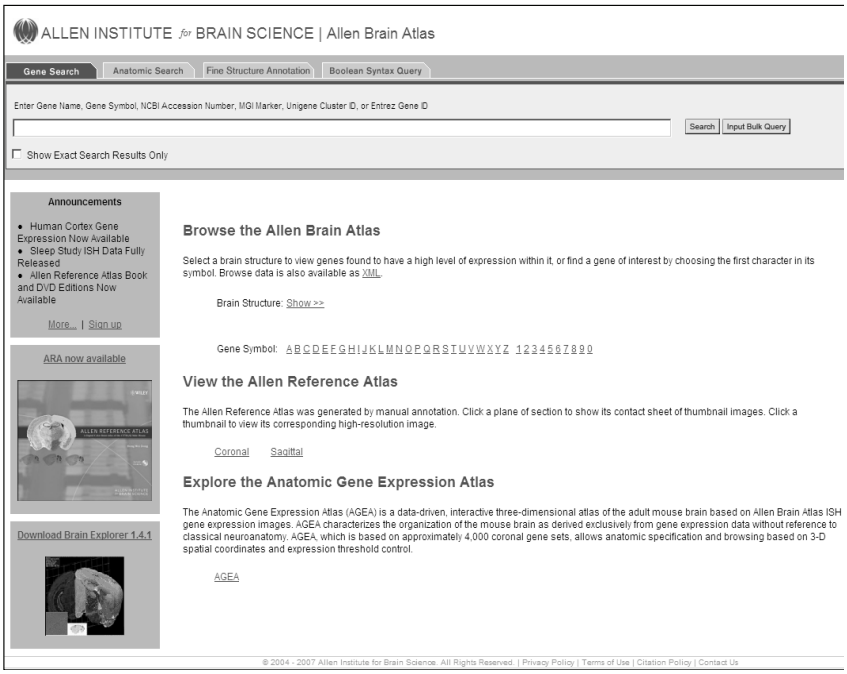

Figure 2. Screen dump of the Allen Brain Atlas

Table 4. Supply- and demand-side attribute values of Allen Brain Atlas

	Dimension	Attribute	Attribute values
S1	Access	Restriction	*none*
		Access media	*on-or-offline*
		Granularity	*data-item*
		Functionality	*medium*
		Software-requirements	*server-resident-proprietary*
S2	Scale, dynamism, coverage and completeness	Scale	*large*
		Dynamism of discipline	*dynamic*
		Dynamism (of database)	*dynamic*
		Temporal-depth	*current-only*
		Coverage	*narrow*
		Completeness	*medium*
		Collection-strategy	*active*
		Processing	*intensive*
		Validation	*intensive*
		Timeliness	*medium*
S3	Disciplinary usage	Cross-discipline	*somewhat*
		Disciplines	*genomics, neuro-anatomy, molecular biology*
		Level of user support	*low*
S4	Interface	User-interface	*graphical-selection*
		Programmable-interfaces	*no*
S5	Interoperability	Self-describing data	*somewhat*
		Semantic transparency	*yes*
		Linkage-to-other-collections	*yes*
		Use of semantic standards	*somewhat*
		Cross-domain semantic cross-walks	*no*
		Programmable interfaces	*non-existent*
S6	Ownership, funding, governance, management and contributors	Reputation	*medium*
		Involvement	*low*
		Accessibility	*??*
		Funding-level	*high*
		Funding-reliability	*medium*
		Governance-quality	*??*
		Sustainability	*medium-term*
S7	s7. Attribution & IP	Attribution completeness	*n/a (self-generated data)*
		Attribution accuracy	*n/a*
		Attribution granularity	*n/a*
		Licensing, registration, agreements with owners	*n/a*
		End-user licensing	*n/a*
		Redaction/anomalisation of data	*n/a*

	Dimension	Attribute	Attribute values
D1	Research methodology, funding and stakeholder requirements	Required-access-granularity	*query-result*
		Required-metadata	*high*
		Required-access-to-models	*low*
		Methods	*<method designations>*
		Publication/distribution requirements	*??*

	Dimension	Attribute	Attribute values
D2	Discovery methods	Search-engines	*??*
		Discovery metadata	*??*
		Other discovery resources	*??*
D3	Query style	Expressivity	*medium*
		Desired-interface	*graphical-selection*
		Required-programmable-access	*low*
D4	Federation	Need-to-federate	*low*
		Required-metadata-support	*low*
D5	Cross-disciplinary usage	Cross-disciplinary-usage	*medium*
		Required-metadata-support	*medium*
D6	Timeliness and temporal access	Required-recency	*medium*
		Required-timestamp-granularity	*low*
		Desired-update-method	*asynchronous*
		Required-temporal-access	*current-only*

5.2 PBDsum

PDBsum provides an overview of every macromolecular structure deposited in the Protein Data Bank (PDB), giving schematic diagrams of the molecules in each structure and of the interactions between them. The database is hosted by the European Bioinformatics Institute, which is part of the European Molecular Biology Laboratory in Cambridgeshire, United Kingdom.

Managed by: European Bioinformatics Institute

Covers: 51,171 entries

URL: http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/

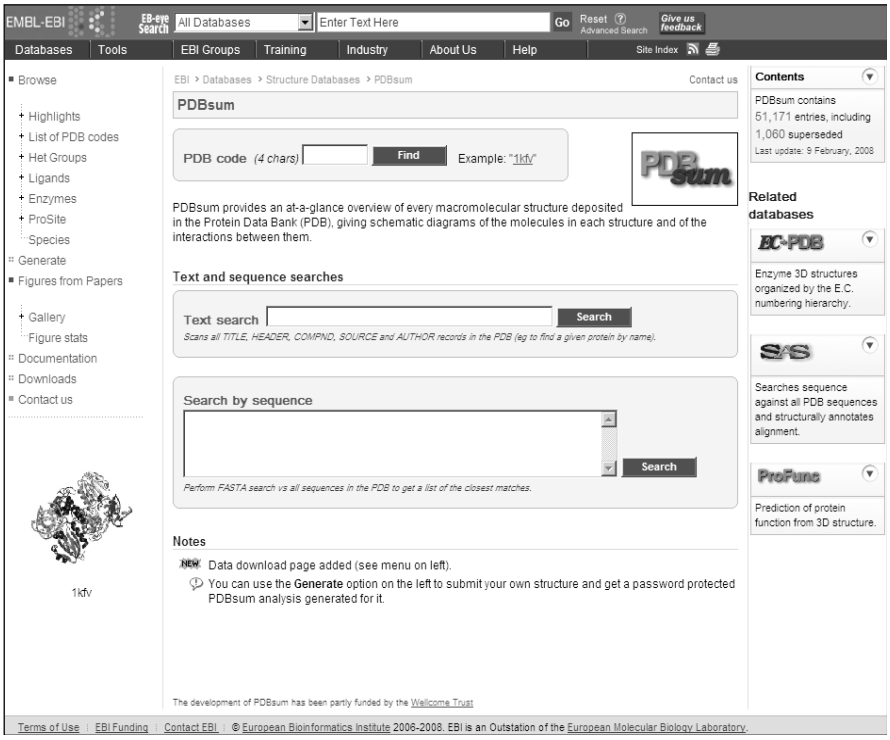

Figure 3. Screen dump of PBDsum

Table 5. Supply- and demand-side attribute values of PBDsum

	Dimension	Attribute	Attribute values
S1	Access	Restriction	*none*
		Access media	*online-only*
		Granularity	*data-item*
		Functionality	*medium*
		Software-requirements	*free-download*
S2	Scale, dynamism, coverage and completeness	Scale	*large*
		Dynamism of discipline	*dynamic*
		Dynamism (of database)	*dynamic*
		Temporal-depth	*versioned*
		Coverage	*broad*
		Completeness	*high*
		Collection-strategy	*passive*
		Processing	*significant*
		Validation	*significant*
		Timeliness	*medium*
S3	Disciplinary usage	Cross-discipline	*somewhat*
		Disciplines	*molecular biology*
		Level of user support	*high*
S4	Interface	User-interface	*text-query*
		Programmable-interfaces	*no*
S5	Interoperability	Self-describing data	*somewhat*
		Semantic transparency	*somewhat*
		Linkage-to-other-collections	*yes*
		Use of semantic standards	*somewhat*
		Cross-domain semantic cross-walks	*somewhat*
		Programmable interfaces	*non-existent*
S6	Ownership, funding, governance, management and contributors	Reputation	*high*
		Involvement	*high*
		Accessibility	*low*
		Funding-level	*medium*
		Funding-reliability	*high*
		Governance-quality	*high*
		Sustainability	*indefinite*
S7	Attribution & IP	Attribution completeness	*n/a*
		Attribution accuracy	*n/a*
		Attribution granularity	*n/a*
		Licensing, registration, agreements with owners	*??*
		End-user licensing	*n/a*
		Redaction/anomalisation of data	*n/a*

	Dimension	Attribute	Attribute values
D1	Research methodology, funding and stakeholder requirements	Required-access-granularity	*data-item*
		Required-metadata	*medium*
		Required-access-to-models	*low*
		Methods	*??*
		Publication/distribution requirements	*??*
D2	Discovery methods	Search-engines	??
		Discovery metadata	??
		Other discovery resources	??
D3	Query style	Expressivity	*high*
		Desired-interface	*text-query*

	Dimension	Attribute	Attribute values
		Required-programmable-access	*??*
D4	Federation	Need-to-federate	*low*
		Required-metadata-support	*low*
D5	Cross-disciplinary usage	Cross-disciplinary-usage	*medium*
		Required-metadata-support	*medium*
D6	Timeliness and temporal access	Required-recency	*high*
		Required-timestamp-granularity	*low*
		Desired-update-method	*asynchronous*
		Required-temporal-access	*current-only*

5.3 PubChem

PubChem provides information on the biological activities of small molecules. PubChem includes substance information, compound structures, and BioActivity data in three primary databases: 1) Pcsubstance contains more than 38 million records of substance records; 2) Pccompound contains more than 18 million unique structures; and 3) PCBioAssay contains more than 800 BioAssays, with each BioAssay containing a various number of data points.

Managed by: National Center for Biotechnology Information, National Institutes of Health

Covers: More than 38 million records of substance records, 18 million structures, and 800 bioAssays

URL: http://pubchem.ncbi.nlm.nih.gov/

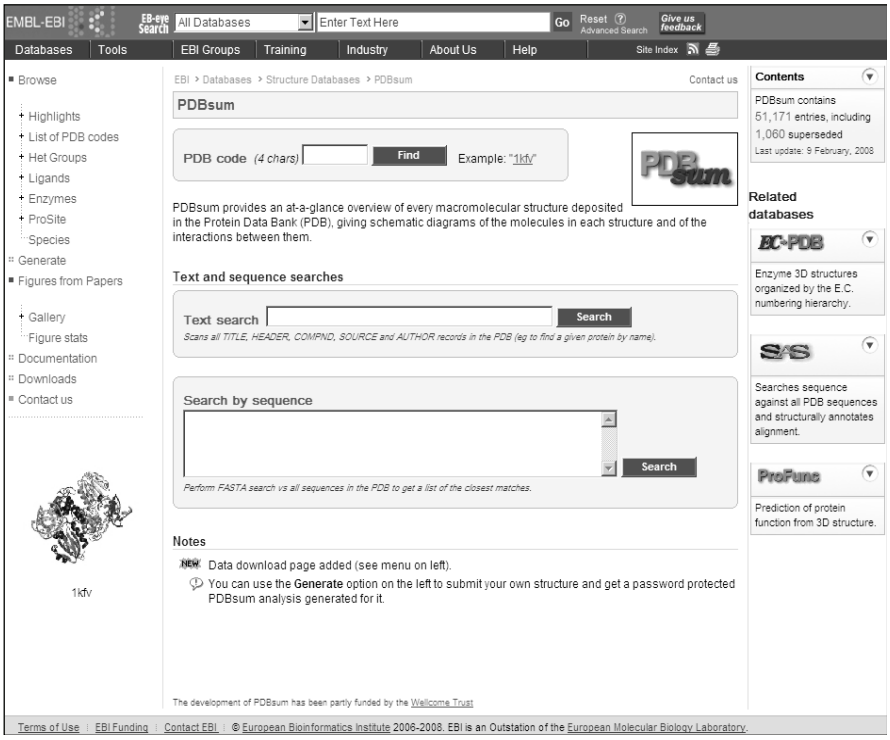

Figure 4. Screen dump of PubChem

Table 6. Supply- and demand-side attribute values of PubChem

	Dimension	Attribute	Attribute values
S1	Access	Restriction	*none*
		Access media	*online-only*
		Granularity	*data-item*
		Functionality	*high*
		Software-requirements	*server-resident-proprietary*
S2	Scale, dynamism, coverage and completeness	Scale	*large*
		Dynamism of discipline	*dynamic*
		Dynamism (of database)	*dynamic*
		Temporal-depth	*current-only*
		Coverage	*broad*
		Completeness	*high*
		Collection-strategy	*passive*
		Processing	*minimal*
		Validation	*minimal ??*
		Timeliness	*medium*
S3	Disciplinary usage	Cross-discipline	*somewhat*
		Disciplines	*chemistry, bio-chemistry, medicine*
		Level of user support	*low*
S4	Interface	User-interface	*graphics-input*
		Programmable-interfaces	*no*
S5	Interoperability	Self-describing data	*yes*
		Semantic transparency	*yes*
		Linkage-to-other-collections	*yes*
		Use of semantic standards	*yes*
		Cross-domain semantic cross-walks	*no*
		Programmable interfaces	*non-existent*
S6	Ownership, funding, governance, management and contributors	Reputation	*high*
		Involvement	*high*
		Accessibility	*high*
		Funding-level	*medium*
		Funding-reliability	*medium*
		Governance-quality	*high*
		Sustainability	*medium-term*
S7	Attribution & IP	Attribution completeness	*high*
		Attribution accuracy	*??*
		Attribution granularity	*high*
		Licensing, registration, agreements with owners	*complete*
		End-user licensing	*n/a*
		Redaction/anomalisation of data	*n/a*

	Dimension	Attribute	Attribute values
D1	Research methodology, funding and stakeholder requirements	Required-access-granularity	*data-item*
		Required-metadata	*medium*
		Required-access-to-models	*low*
		Methods	*??*
		Publication/distribution requirements	*??*

	Dimension	Attribute	Attribute values
D2	Discovery methods	Search-engines	*??*
		Discovery metadata	*??*
		Other discovery resources	*??*
D3	Query style	Expressivity	*high*
		Desired-interface	*graphics-input*
		Required-programmable-access	*low*
D4	Federation	Need-to-federate	*medium*
		Required-metadata-support	*medium*
D5	Cross-disciplinary usage	Cross-disciplinary-usage	*medium*
		Required-metadata-support	*medium*
D6	Timeliness and temporal access	Required-recency	*high*
		Required-timestamp-granularity	*low*
		Desired-update-method	*asynchronous*
		Required-temporal-access	*current-only*

5.4 FishBase

FishBase is a relational database with information to cater to different professionals such as research scientists, fisheries managers, and zoologists. This online resource claims to contain practically all fish species known to science. It was developed at the WorldFish Center in collaboration with the Food and Agriculture Organisation of the United Nations.

Managed by: A consortium of seven research institutions, including the WorldFish Center Food and Agriculture and the Organisation of the United Nations

Covers: 30100 Species, 262300 Common names, 45000 Pictures and 40600 References

URL: http://www.fishbase.org/home.htm

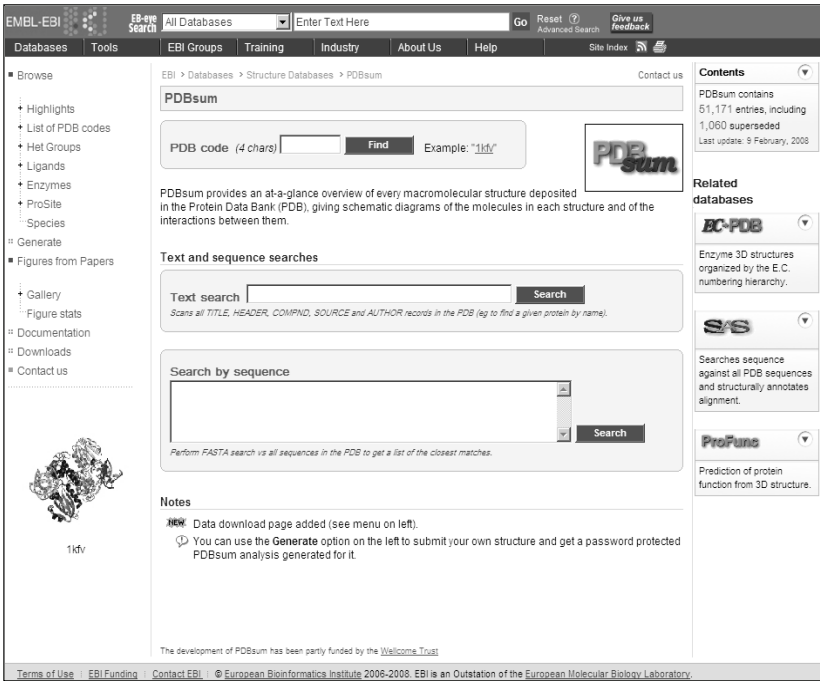

Figure 5. Screen dump of FishBase

Table 7. Supply- and demand-side attribute values of FishBase

	Dimension	Attribute	Attribute values
S1	Access	Restriction	*none*
		Access media	*on-or-offline*
		Granularity	*query-result*
		Functionality	*medium*
		Software-requirements	*server-resident-proprietary*
S2	Scale, dynamism, coverage and completeness	Scale	*large*
		Dynamism of discipline	*static*
		Dynamism (of database)	*static*
		Temporal-depth	*current-only*
		Coverage	*broad*
		Completeness	*high*
		Collection-strategy	*??*
		Processing	*mimimal*
		Validation	*??*
		Timeliness	*low*
S3	Disciplinary usage	Cross-discipline	*somewhat*
		Disciplines	*ichthyology, zoology, ecology, fisheries management*
		Level of user support	*high*
S4	Interface	User-interface	*graphical-selection*
		Programmable-interfaces	*no*
S5	Interoperability	Self-describing data	*no*
		Semantic transparency	*no*
		Linkage-to-other-collections	*somewhat*
		Use of semantic standards	*somewhat*
		Cross-domain semantic cross-walks	*no*
		Programmable interfaces	*non-existent*
S6	Ownership, funding, governance, management and contributors	Reputation	*high*
		Involvement	*medium*
		Accessibility	*??*
		Funding-level	*medium*
		Funding-reliability	*high*
		Governance-quality	*high*
		Sustainability	*indefinite*
S7	Attribution & IP	Attribution completeness	*high*
		Attribution accuracy	*??*
		Attribution granularity	*high*
		Licensing, registration, agreements with owners	*??*
		End-user licensing	*minimal*
		Redaction/anomalisation of data	*n/a*

	Dimension	Attribute	Attribute values
D1	Research methodology, funding and stakeholder requirements	Required-access-granularity	*data-item*
		Required-metadata	*low*
		Required-access-to-models	*low*
		Methods	*??*
		Publication/distribution requirements	*??*
D2	Discovery methods	Search-engines	??
		Discovery metadata	??
		Other discovery resources	??
D3	Query style	Expressivity	*medium*
		Desired-interface	*text-query*
		Required-programmable-access	*low*
D4	Federation	Need-to-federate	*low*
		Required-metadata-support	*low*
D5	Cross-disciplinary usage	Cross-disciplinary-usage	*low*
		Required-metadata-support	*low*
D6	Timeliness and temporal access	Required-recency	*low*
		Required-timestamp-granularity	*low*
		Desired-update-method	*asynchronous*
		Required-temporal-access	*versioned*

5.5 Combined mapping of sample of collections

The attribute values in the following table are compiled by either ends of the possible range gathered from the four databases under consideration. With the caveat that it is based on a very small sample size, they represent the range of plausible attribute values of supply- and demand-side dimensions for reference data collections in the fields of STM.

Table 8. Combined supply- and demand-side attribute values of four sample collections.

	Dimension	Attribute	Attribute values
S1	Access	Restriction	*none*
		Access media	*on-or-offline*
		Granularity	*data-item/query-result*
		Functionality	*medium/high*
		Software-requirements	*server-resident-proprietary/free-download*
S2	Scale, dynamism, coverage and completeness	Scale	*large*
		Dynamism of discipline	*dynamic*
		Dynamism (of database)	*dynamic*
		Temporal-depth	*current-only*
		Coverage	*broad/narrow*
		Completeness	*high/medium*
		Collection-strategy	*passive/active*
		Processing	*minimal/significant/intensive*
		Validation	*significant/intensive*
		Timeliness	*low/medium*
S3	Disciplinary usage	Cross-discipline	*somewhat*
		Disciplines	*<varied>*
		Level of user support	*low-high*
S4	Interface	User-interface	*<varied>*
		Programmable-interfaces	*no*
S5	Interoperability	Self-describing data	*no/somewhat*
		Semantic transparency	*<varied>*

	Dimension	Attribute	Attribute values
		Linkage-to-other-collections	*somewhat/yes*
		Use of semantic standards	*somewhat/yes*
		Cross-domain semantic cross-walks	*no/somewhat*
		Programmable interfaces	*non-existent*
S6	Ownership, funding, governance, management and contributors	Reputation	*medium/high*
		Involvement	*<varied>*
		Accessibility	*<varied>*
		Funding-level	*medium/high*
		Funding-reliability	*medium/high*
		Governance-quality	*<varied>*
		Sustainability	*medium-term/indefinite*
S7	Attribution & IP	Attribution completeness	*<varied>*
		Attribution accuracy	*??*
		Attribution granularity	*<varied>*
		Licensing, registration, agreements with owners	*<varied>*
		End-user licensing	*n/a*
		Redaction/anomalisation of data	*n/a*

Table 7 (cont'd). Combined supply- and demand-side attribute values of four sample collections.

	Dimension	Attribute	Attribute values
D1	Research methodology, funding and stakeholder requirements	Required-access-granularity	*data-item/query-result*
		Required-metadata	*<varied>*
		Required-access-to-models	*low*
		Methods	*??*
		Publication/distribution requirements	*??*
D2	Discovery methods	Search-engines	??
		Discovery metadata	??
		Other discovery resources	??
D3	Query style	Expressivity	*medium/high*
		Desired-interface	*<varied>*
		Required-programmable-access	*low*
D4	Federation	Need-to-federate	*low/medium*
		Required-metadata-support	*low/medium*
D5	Cross-disciplinary usage	Cross-disciplinary-usage	*low/medium*
		Required-metadata-support	*low/medium*
D6	Timeliness and temporal access	Required-recency	*<varied>*
		Required-timestamp-granularity	*low*
		Desired-update-method	*asynchronous*
		Required-temporal-access	*current-only/versioned*

CHAPTER 6 Alternatives for the national library

This chapter synthesises the findings from previous chapters and the consequences for the BL. Previous chapters show that there is a variety of feasible supply- and demand-side characteristics of potential reference data collections. The theoretical scope of this variation is spanned by a rather unwieldy matrix consisting of the number of supply-side attributes on the one hand and the number of demand-side attributes on the other. Unless a significant number of these attributes turn out to be irrelevant, it therefore seems best to allow this matrix to remain merely conceptual.

It seems infeasible for a data centre to develop a single data access strategy that comprehensively accommodates the full range of data collection characteristics and user preferences. Therefore, instead of trying to suggest such a comprehensive strategy for the BL, we identify a range of optional approaches that address each or a small set of salient attribute values. The appendix to this report discusses these available optional approaches for the national library in more detail. For each attribute within the dimensions, it discusses one or more options that address the relevant supply- and demand-side issues.

Some options listed in the appendix partially address issues of long-term preservation, archiving, and the provision of future access. The issues and the wide range of options associated with long-term preservation are so complicated in itself that we have focused on options for enabling access.[7] The question of whether or not the BL should take on these tasks for a database is broader than the questions addressed by most of the options listed. So although some specific preservation options, for example archiving or dark archiving, are listed for several specific attributes, they may also be appropriate in other cases.

The purpose of characterising databases as we have done is to attempt to generate coherent sets of options for the BL with respect to specific classes of databases, where these options are derived from the nature of the databases in each such class. Ideally, the attribute values of each individual database would be mapped to options, but due to the somewhat tentative nature of the values we have obtained for these attributes, we believe it is more useful to map this abstracted cluster to options rather than mapping each individual database, which might give the impression of greater accuracy and reliability than our limited analysis warrants.

[7] For a more detailed discussion of the strategic issues related to long-term preservation for a national library, see for example: Hoorens *et al.* (2007).

Each point-cluster characterises a class of databases that share certain key attribute values. The set of options listed in each cluster suggests how the BL might approach databases in this class. In this section we have identified three such clusters:

1. A "neutral" cluster: based on the combined attribute values of the sample of data collections;
2. A "demanding" cluster: based on a complex, demanding set of requirements combined with relatively minimum support by the database itself;
3. An "undemanding" cluster: based on a simple, undemanding set of requirements combined with relatively good support by the database itself.

These three clusters, however, should be considered indicative rather than definitive.

The following discussion uses the terms "gateway" and "transparent" access, as defined above. To reiterate: gateway access means providing access to a data collection via some alternative site (such as the BL) other than the collection's home site, without reproducing or re-hosting the collection or its fundamental access mechanisms. A gateway may, however, provide additional access or query mechanisms or additional curatorial or explanatory services beyond those offered by the collection's home site. In contrast, transparent access means providing alternative access to a data collection without reproducing or re-hosting the collection or its access mechanisms, and without providing any additional access or query mechanisms or additional curatorial or explanatory services beyond those offered by the collection's home site.

6.1 Point-cluster 1: "Neutral" case

The first cluster of options addresses the issues arising from the sample of four databases that were investigated: Allen Brain Atlas, PubChem, FishBase and PDBsum. Assuming that these data collections are a good representation of the available high quality reference data collections, this cluster can be considered as a neutral approach, anticipating the required effort, for the role of a national library. In this cluster, we would recommend considering providing transparent access to the data collections. This means providing alternative access to a data collection without reproducing or re-hosting the collection or its access mechanisms, and without providing any additional access or query mechanisms or additional curatorial or explanatory services beyond those offered by the collection's home site. The options for each attribute in this cluster are illustrated in Table 9.

6.2 Point-cluster 2: "Demanding" case

The second cluster of options represents a case in which the national library addresses a complex, demanding set of requirements combined with relatively minimum support by the database itself. In this cluster, we would recommend considering providing gateway access to the data collections. This would involve facilitating access to a data collection via an alternative site, for example that of the BL, other than the collection's home site, without reproducing or re-hosting the collection or its fundamental access mechanisms. The gateway could provide additional access or query mechanisms or additional curatorial

or explanatory services beyond those offered by the collection's home site. The options for each attribute in this cluster are illustrated in Table 10.

6.3 Point-cluster 3: "Undemanding" case

The third cluster represents a case in which the role of the national library addresses a simple, undemanding set of requirements, combined with relatively good support by the database itself. The data collections require little in the way of access restriction, and the supporting mechanisms are relatively simple. In this cluster, we would recommend considering providing transparent access to the data collections (see above). The options for each attribute in this cluster are illustrated in Table 11.

Table 9. Salient point-cluster 1: Neutral case

	Dimension		Attribute	Attribute values	Options
S1	Access	*	Restriction	*none*	1
			Access media	*on-or-offline*	3
			Granularity	*data-item*	1
		*	Functionality	*medium*	1
		*	Software-requirements	*server-resident-proprietary*	1
S2	Scale, dynamism, coverage and completeness	*	Scale	*large*	2
		*	Dynamism of discipline	*dynamic*	2
		*	Dynamism (of database)	*dynamic*	2
			Temporal-depth	*current-only*	1
			Coverage	*broad*	1
		*	Completeness	*high*	1
			Collection-strategy	--	
			Processing	--	
		*	Validation	*significant/intensive*	1
		*	Timeliness	*medium*	1
S3	Disciplinary usage		Cross-discipline	--	
			Disciplines	--	
			Level of user support	*medium*	2
S4	Interface	*	User-interface	*<simple>*	1
		*	Programmable-interfaces	*no*	1
S5	Interoperability		Self-describing data	*somewhat*	
			Semantic transparency	*<varied>*	
			Linkage-to-other-collections	--	
			Use of semantic standards	*somewhat/yes*	1
			Cross-domain semantic cross-walks	*no/somewhat*	1
		*	Programmable interfaces	*non-existent*	
S6	Ownership, funding, governance, management and contributors	*	Reputation	*medium/high*	1
			Involvement	--	
		*	Accessibility	*??*	2
			Funding-level	*medium/high*	
		*	Funding-reliability	*medium/high*	1
		*	Governance-quality	*??*	???
		*	Sustainability	*medium-term/indefinite*	1
S7	Attribution & IP	*	Attribution completeness	*<varied>*	???
		*	Attribution accuracy	*??*	???
		*	Attribution granularity	--	
			Licensing, registration,	*<varied>*	???

	Dimension		Attribute	Attribute values	Options
			agreements with owners		
			End-user licensing	*n/a*	1
			Redaction/anomalisation of data	*n/a*	1

	Dimension		Attribute	Attribute values	Options
D1	Research methodology, funding and stakeholder requirements	*	Required-access-granularity	*data-item/query-result*	1
		*	Required-metadata	*<varied>*	2
		*	Required-access-to-models	*low*	1
			Methods	--	
			Publication/distribution requirements	*??*	???
D2	Discovery methods		Search-engines	*??*	???
			Discovery metadata	*??*	???
			Other discovery resources	*??*	???
D3	Query style	*	Expressivity	*medium/high*	1
		*	Desired-interface	*<varied>*	1
		*	Required-programmable-access	*low*	1
D4	Federation		Need-to-federate	*low/medium*	
		*	Required-metadata-support	*low/medium*	1
D5	Cross-disciplinary usage		Cross-disciplinary-usage	*low/medium*	
		*	Required-metadata-support	*low/medium*	1
D6	Timeliness and temporal access	*	Required-recency	*<varied>*	1
		*	Required-timestamp-granularity	*low*	2
		*	Desired-update-method	*asynchronous*	2
		*	Required-temporal-access	*current-only/versioned*	1

Note: attributes preceded by "*" are those deemed especially relevant to the BL in deciding on a strategy with respect to a particular database. The numbers in the "Options" column refer to the options outlined in Appendix A.

Table 10. Salient point-cluster 2. "Demanding" case

	Dimension		Attribute	Attribute values	Options
S1	Access	*	Restriction	none	2+D
			Access media	online-only/ on-or-offline	2
			Granularity	data-item	1
		*	Functionality	medium	3
		*	Software-requirements	server-resident-proprietary	3
S2	Scale, dynamism, coverage and completeness	*	Scale	large	3
		*	Dynamism of discipline	dynamic	3
		*	Dynamism (of database)	dynamic	2
			Temporal-depth	current-only	1
			Coverage	broad	
		*	Completeness	high	
			Collection-strategy	--	
			Processing	--	
		*	Validation	significant/intensive	2
		*	Timeliness	medium	3
S3	Disciplinary usage		Cross-discipline	--	1
			Disciplines	--	
			Level of user support	low	1
S4	Interface	*	User-interface	<simple>	2
		*	Programmable-interfaces	no	2
S5	Interoperability		Self-describing data	somewhat	
			Semantic transparency	<varied>	
			Linkage-to-other-collections	--	2
			Use of semantic standards	no	2
			Cross-domain semantic cross-walks	no	2
		*	Programmable interfaces	non-existent	2
S6	Ownership, funding, governance, management and contributors	*	Reputation	medium/high	1
			Involvement	--	
		*	Accessibility	??	2
			Funding-level	medium/high	
		*	Funding-reliability	medium/high	2+A,D
		*	Governance-quality	??	3
		*	Sustainability	medium-term/indefinite	2
S7	Attribution & IP	*	Attribution completeness	<varied>	2
		*	Attribution accuracy	??	2
		*	Attribution granularity	--	
			Licensing, registration, agreements with owners	partial	2
			End-user licensing	partial	2
			Redaction/anomalisation of data	as-needed	2

	Dimension		Attribute	Attribute values	Options
D1	Research methodology, funding and stakeholder requirements	*	Required-access-granularity	data-item/query-result	1
		*	Required-metadata	medium	2
		*	Required-access-to-models	low	1
			Methods	--	
			Publication/distribution requirements	--	2
D2	Discovery methods		Search-engines	specialised	2
			Discovery metadata	specialised	2
			Other discovery resources	idiosyncratic	2

	Dimension		Attribute	Attribute values	Options
D3	Query style	*	Expressivity	medium/high	2
		*	Desired-interface	<varied>	3
		*	Required-programmable-access	low	2
D4	Federation		Need-to-federate	low/medium	2
		*	Required-metadata-support	low/medium	1
D5	Cross-disciplinary usage		Cross-disciplinary-usage	low/medium	1
		*	Required-metadata-support	low/medium	1
D6	Timeliness and temporal access	*	Required-recency	<varied>	4
		*	Required-timestamp-granularity	low	1
		*	Desired-update-method	asynchronous	1
		*	Required-temporal-access	current-only/versioned	1

Note: attributes preceded by "*" are those deemed especially relevant to the BL in deciding on a strategy with respect to a particular database. The numbers in the "Options" column refer to the options outlined in Appendix A.

Table 11. Salient point-cluster 3. "Undemanding" case

	Dimension		Attribute	Attribute values	options
S1	Access	*	Restriction	none	1
			Access media	online-only/ on-or-offline	2
			Granularity	data-item	
		*	Functionality	medium	1
		*	Software-requirements	server-resident-proprietary	1
S2	Scale, dynamism, coverage and completeness	*	Scale	large	1
		*	Dynamism of discipline	dynamic	1
		*	Dynamism (of database)	dynamic	1
			Temporal-depth	current-only	
			Coverage	broad	
		*	Completeness	high	1
			Collection-strategy	--	
			Processing	--	
		*	Validation	significant/intensive	1
		*	Timeliness	medium	1
S3	Disciplinary usage		Cross-discipline	--	
			Disciplines	--	
			Level of user support	high	2
S4	Interface	*	User-interface	<simple>	1
		*	Programmable-interfaces	no	1
S5	Interoperability		Self-describing data	somewhat	
			Semantic transparency	<varied>	
			Linkage-to-other-collections	--	
			Use of semantic standards	yes	1
			Cross-domain semantic cross-walks	yes	1
		*	Programmable interfaces	non-existent	
S6	Ownership, funding, governance, management and contributors	*	Reputation	medium/high	2
			Involvement	--	
		*	Accessibility	??	1
			Funding-level	medium/high	
		*	Funding-reliability	medium/high	1
		*	Governance-quality	??	2
		*	Sustainability	medium-term/indefinite	1
S7	Attribution & IP	*	Attribution completeness	<varied>	3
		*	Attribution accuracy	??	3
		*	Attribution granularity	--	
			Licensing, registration, agreements with owners	n/a	1
			End-user licensing	n/a	1
			Redaction/anomalisation of data	n/a	1

	Dimension		Attribute	Attribute values	Options
D1	Research methodology, funding and stakeholder requirements	*	Required-access-granularity	data-item/query-result	1
		*	Required-metadata	medium	1
		*	Required-access-to-models	low	1
			Methods	--	
			Publication/distribution requirements	n/a	1

	Dimension		Attribute	Attribute values	Options
D2	Discovery methods		Search-engines	generic	
			Discovery metadata	generic	
			Other discovery resources	none	
D3	Query style	*	Expressivity	medium/high	1
		*	Desired-interface	<varied>	1
		*	Required-programmable-access	low	1
D4	Federation		Need-to-federate	low/medium	
		*	Required-metadata-support	low/medium	1
D5	Cross-disciplinary usage		Cross-disciplinary-usage	low/medium	
		*	Required-metadata-support	low/medium	1
D6	Timeliness and temporal access	*	Required-recency	<varied>	1
		*	Required-timestamp-granularity	low	2
		*	Desired-update-method	asynchronous	2
		*	Required-temporal-access	current-only/versioned	1

Note: attributes preceded by "*" are those deemed especially relevant to the BL in deciding on a strategy with respect to a particular database. The numbers in the "Options" column refer to the options outlined in Appendix A.

CHAPTER 7 Lessons from this approach

The set of options listed in each cluster presented in the previous chapter offers a potential strategy that the BL may apply to databases in a given class. However, these options are only a starting point. The BL's strategy with respect to any given database should be decided on the basis of an overall assessment of the importance and uniqueness of that database, its relevance to the BL's policies with regard to STM data, and the BL's assessment of the degree to which users of the database would benefit from having the BL apply its own curatorial, preservation, or access resources to the database.[8] Our analytic approach, required an average of about a person-day for each database (with significant variance). The basic analysis of each site did not take more than half a day, but when including the following of all of the links to background information, related sites, ownership and funding organisations, etc, the total time spent adds to about a person-day. We believe that devoting more time to this method would be diminishingly productive, since most databases provide relatively little contextual information on their websites. Notably absent in most cases is much if any information about data collection or generation methods, validation, metadata, funding, governance, usage methods and usage patterns.

Our approach provided reasonable information for most supply-side attributes, though details of ownership and funding (accessibility of owners, owner reputation, reliability of funding, etc) could in many cases only be inferred by this method. Demand-side attributes were even harder to obtain by means of website analysis. In some cases, links to user groups or searches for communities that seemed likely to utilise a given database provided hints about research methods, preferred query styles, the degree of required federation or cross-disciplinary usage and the need for temporal access; but our values for most of these attributes are essentially informed guesses, which need to be validated and revised based on future demand-side analysis.

The lack of demand-side input also limits the validity of our values for some supply-side attributes, such as the reputation of the database's ownership and management, the quality of its data and metadata, its completeness, timeliness, etc.

In light of the limited contextual information that is available from the websites of even the best supported databases, we recommend a more in-depth examination that employs direct contact with database administrators, parent organisations, data processing

[8] The example of Intute (See Chapter 1) illustrates how these issues can be addressed. For more details, see: Burgess (2006).

managers, discipline-based organisations whose members use the database, and user communities. This should help fill in the supply-side attributes for a database as well as providing demand-side attributes, whose values were supplied largely by assumptions in the current study. We estimate that an average of 3 to 5 person-days of effort devoted to each database should be sufficient to provide more reliable and complete values for the attributes we have identified. Our suggested procedure for performing this more extensive research would be:

1. Spend about a person-day browsing the database's website to provide context, follow promising links, fill in obvious attributes, and generate a list of missing attribute values.
2. Contact database owners, managers, administrators, data processing and validation personnel, and others, as appropriate, to fill in missing supply-side attributes and provide links to demand-side groups, such as discipline-based organisations, user communities and forums, etc.

Conduct interviews with different demand-side groups to fill in demand-side attributes and to validate supply-side attributes such as the reputation and accessibility of the database's ownership and management, the quality of its data and metadata, its accessibility and usability, its completeness and timeliness, etc. Additionally, a survey of a sample of users would provide a more quantitative evidence base for the demand-side characteristics of data collections relevant to the BL

REFERENCES

Reference list

Burgess J (2006) Intute Collection Development Framework and Policy. Version 1.4 Approved by the Intute Policy and Strategy Forum 12/10/06. Available at: http://www.intute.ac.uk/policy.html

European Commission (2007) Communication from the Commission to the European Parliament, the Council and the European Economic and Social Committee on Scientific Information in the Digital Age: Access, Dissemination and Preservation (Commission of the European Communities. February 14, 2007. Available at: http://ec.europa.eu/information_society/activities/digital_libraries/doc/scientific_information/communication_en.pdf

Hoorens, Stijn, Jeff Rothenberg, Constantijn van Oranje, Martin van der Mandele and Ruth Levitt (2007) Addressing the uncertain future of preserving the past: Towards a robust strategy for digital archiving and preservation. RAND Europe TR-510-KB, Cambridge. Available at: http://www.rand.org/pubs/technical_reports/TR510/

Lyon, Liz (2007) Dealing with Data: Roles, Rights, Responsibilities and Relationships. Consultancy Report, UKOLN, University of Bath. 19th June 2007. Available at: http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf

National Science foundation (2005) Long-lived digital data collections: enabling research and education in the 21st century. September, 2005. Report of the National Science Board. Available at: http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf

National Science Foundation (2007) NSF's Cyberinfrastructure Vision for 21st Century Discovery. Report of the National Science Board. Available at: http://www.nsf.gov/pubs/2007/nsf0728/index.jsp

OECD (2004) Science, Technology and Innovation for the 21st Century. Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 29-30 January 2004 - Final Communiqué (Organisation for Economic Co-operation and Development. January 30, 2004. Available at: http://www.oecd.org/document/15/0,2340,en_21571361_21590465_25998799_1_1_1_1,00.html

RCUK (2005) RCUK position on issue of improved access to research outputs. Available at: http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/documents/2006statement.pdf

APPENDIX

Appendix A. Options for BL

S1. Access

Restriction

The less restricted a database is, the easier it would be for the BL to re-host or provide indirect access to it. However, this would also make it easier for users to access the database directly, so access per se would not contribute motivation in such cases. Databases having restricted access might therefore be a more attractive target for the BL, assuming the BL can work out some arrangement for providing wider access without undermining the database's business and governance models.

If a database restricts access because of concerns about possible misuse or misinterpretation of its data by unregistered users, the BL might be able to enforce appropriate usage restrictions, serving as a "gateway" that would alleviate this concern while providing access to a broader community.

If a database restricts access as part of its business model (for example, to obtain payment via membership or usage fees), the BL might be able to negotiate an arrangement that would provide wider access while providing sufficient monetary support to the database.

Option 1: Provide transparent access to the database
Option 2: Serve as a gateway to provide restricted access
Option 3: Re-host the database
Option-D: Provide a "dark archive" for the database

Access media

If a database is available in offline form, such as DVD, the BL may or may not decide to provide an alternate publication source for such media. Similarly, if a database is available in bulk download from, the BL may or may not offer such download capability itself.

Option 1: Serve as an alternative offline and/or download data publisher
Option 2: Transparently pass requests for offline data to original data publisher
Option 3: Provide online-only access to data

Granularity

If the BL provides transparent access to a database, the granularity of access should not be a significant factor, except in terms of the bandwidth that may be required to pass through the BL to the database itself. If the BL re-hosts a database, however, granularity of access will impact the kind of interface the BL must provide, as well as the bandwidth needed to supply data to users.

Option 1: Ignore granularity as being an insignificant factor
Option 2: Consider granularity when performing sizing/load estimates

Functionality

The access functionality provided by a database goes beyond its restriction and granularity. Some databases provide only literal data values, whereas others provide graphical or interactive visualisation tools. Some enable searching for individual items, whereas others allow queries that return complex sets of related data. Some allow browsing taxonomic trees or graphical representations, whereas others allow searching by means of formal notation or graphical input (for example, sketching molecular structures). Because functionality can be so varied, it is difficult to characterise it beyond giving it an ordinal value, such as low, medium or high. Nevertheless, one of the key motivations for the BL's re-hosting or acting as a gateway to a database may be to provide enhanced (or simplified) access to its data.

Option 1: Rely on the database's native access functionality
Option 2: Reproduce the database's native access functionality
Option 3: Provide reduced, simplified access functionality
Option 4: Provide enhanced access functionality

Software-requirements:

If the BL re-hosts a database or serves as a gateway to it, then the BL may have to bypass, obtain, recreate, utilise, or interface to whatever software the database provides for access. Even if this software is generic, open source, or freely downloadable, the BL may have to maintain and support its own version of it as well as developing in-house expertise in its use. If the software is proprietary or obscure, the BL may need to enter into contractual agreements with the owner or developer of the software or may have to recreate it or replace it, at considerable expense. If the BL serves as a gateway or provides enhanced access beyond that provided by the database itself, it may need to develop or obtain its own specialised software for this purpose and to maintain and support that software for its users.

Option 1: Rely on free software provided by the database
Option 2: Arrange to use proprietary software hosted by the database
Option 3: Obtain or recreate necessary software

S2. Scale, dynamism, coverage and completeness

Scale

If the BL provides transparent access to a database, its scale may not be highly relevant, unless the bandwidth requirements for large dataset downloads must be borne by the BL. The the BL re-hosts or provides a gateway to a database; its scale may impact the BL's storage, processing and support requirements, as well as bandwidth and other access resources (such as the number of servers provided for online access).

Option 1: Assume any added resource requirements can be absorbed
Option 2: Provide transparent access to minimise the BL resource impact
Option 3: Analyse resource impact of the BL's re-hosting or acting as gateway

Dynamism of discipline:

If the disciplines most likely to use a database are dynamic, the structure, content, access mechanisms, and interfaces of the database are more likely to evolve rapidly. This may lead to added cost if the BL re-hosts or provides a gateway to the database, in order to provide evolving functionality and meet a high volume of demand for access. Even if the BL merely provides transparent access to a database, such dynamism may still impact the demand for access and the user community's expectation of rapid response.

Option 1: Assume relatively constant demand and functionality
Option 2: Provide transparent access to minimise the BL resource impact
Option 3: Interact with user groups to meet evolving functionality and volume demands

Dynamism (of database):

Regardless of the dynamism of the disciplines that utilise it, a database may be dynamic in its own right (for example, even an historical database for a closed discipline will be dynamic when it is first being populated with data).

If the BL re-hosts or serves as a gateway to a dynamic database, it may have to provide considerable administrative support, up to and including restructuring or re-implementing the database if its scale or character change. In addition, the tools needed to support large scale data entry and submission may be quite different from those needed to provide access.

Option 1: Assume low levels of data submission and no restructuring
Option 2: Provide transparent access to minimise the BL resource impact
Option 3: Support high levels of data submission and possible restructuring

Temporal-depth

Access to previous values in a database may take several forms, including frozen, historical information, a "snapshot" of current values, multiple versions of data (i.e. previous snapshots), access to specific multiple temporal epochs, or the ability to reconstruct data at

any desired time in the past. If the BL re-hosts or serves as a gateway with enhanced functionality, it may be called upon to provide temporal access beyond that which is available directly from the database itself. This could require check-pointing previous versions of the database, maintaining transaction histories, and enhancing query capabilities to specify temporality.

Option 1: Rely on the database's own temporal storage and access
Option 2: Serve as a gateway offering enhanced temporal queries
Option 3: Host temporal storage and appropriate access mechanisms

Coverage

A database whose coverage is broad is likely to have greater appeal to a wider audience of potential users, which may make it a more attractive target for the BL. On the other hand, databases with narrow coverage may fill a specific niche that the BL may have some reason to support: for example, such a database may disappear or become inaccessible without the BL support. Coverage is relative to a given discipline or community of potential users. In some cases, a discipline or community may be best served by providing a set of narrow coverage databases to produce broad aggregated coverage.

Option 1: Target broad coverage databases for their wide appeal
Option 2: Target narrow coverage databases to ensure their availability
Option 3: Target collections of narrow coverage databases to provide broad aggregated coverage

Completeness

Completeness is relative to coverage and to the dynamism of the disciplines served by a database. Complete databases may in general be more valuable, but incomplete databases may fill important niches, especially when they are more complete than the alternatives. As is the case for coverage, a discipline or community may be best served by providing a set of incomplete databases to produce a more complete set of aggregated data.

Option 1: Target complete databases for their wide appeal
Option 2: Target incomplete databases that fill important niches
Option 3: Target collections of incomplete databases to provide more complete aggregated data

Collection-strategy

Some databases passively await submission of data whereas others actively generate their own data. From the BL's perspective, the main impacts of this distinction would seem to be the credibility and ownership of the resulting data. Data submitted by a range of sources may have varying pedigree, whereas data generated by a database owner will have a single pedigree, which may therefore be easier to evaluate and monitor for quality control purposes.

Similarly, the intellectual property issues surrounding submitted data may be more complex than those for data generated by a database's owner. Whether the BL re-hosts a database, serves as a gateway to it, or provides transparent access to it, the BL must consider its own reputation when putting its "brand" on the resulting data.

Option 1: Rely on the database owner to certify credibility and ownership
Option 2: Track data pedigree and ownership to its ultimate sources

Processing

Related to collection-strategy is the amount of internal processing that a database owner performs on its data. If the owner generates data actively, the processing involved will determine its quality. As above, the BL must consider its own reputation when putting its "brand" on the resulting data.

Presumably, the BL would not become involved in generating data even if it were to re-host a database whose owner generates its own data; however, if the BL serves as a gateway to such a database, it may provide processing of its own, in order to enhance the data or the access functionality that it provides. Even if it does not provide such enhanced capability, the BL may want to examine the processing performed by the database owner in order to be able to add relevant metadata or provide usage support for BL users of the database.

Option 1: Accept any processing performed by the database owner
Option 2: Examine owner processing to add metadata or usage support
Option 3: Perform added processing to enhance data or access functionality

Validation

If a database is populated by submission from parties other than its owner, validity of the data may be assumed or referred to published research results that support the data. A database owner may still perform additional validation on submission, but this is less likely than if the database is generated actively by its owner. As above, the BL must consider its own reputation when putting its "brand" on the resulting data, so it may want to perform validation of its own, though this could be a substantial task for a large or dynamic database.

Option 1: Rely on the database owner to certify the validity of its data
Option 2: Perform additional validation

Timeliness:

The main impact of this attribute for the BL concerns any delay or added latency introduced by re-hosting, serving as a gateway, or even providing transparent access to a database. If a database must be accessed in real-time (such as tsunami early warning data, earthquake data, or astronomical event data), the BL must consider whether its intervention in the access chain may adversely affect timeliness. In some cases, timeliness will not be an issue. In others, any expected delays will be insignificant for the expected

uses of the database. If timeliness is expected to be important, the BL should analyse the probable impact on timeliness of the BL involvement in providing access to the database and use this analysis to help decide whether to re-host, serve as a gateway, provide transparent access, or avoid intervention in the access chain of the given database.

Option 1: Assume that delays will not be a significant factor for users
Option 2: Consider timeliness as part of the BL strategy for the database
Option 3: Monitor and evaluate timeliness when providing access

S3. Disciplinary usage

Cross-discipline:
The main impacts of this attribute for the BL concern the possible need to maintain and provide support for users in multiple disciplines and the possible need to implement access mechanisms or metadata to serve users who are working in multiple fields simultaneously. For example, the BL might provide metadata to enable automated units conversion, semantic mapping, and other translations across discipline boundaries, thereby enhancing the native utility of a database for such interdisciplinary research.

Option 1: Rely on native database cross-discipline support
Option 2: Interact with user groups to support their cross-discipline needs
Option 3: Develop cross-discipline metadata for BL users
Option 4: Develop cross-discipline access mechanisms for BL users

Disciplines:
These are relevant to the BL only to the extent that it may have motivations for supporting specific disciplines rather than others.

Option 1: Target databases regardless of the disciplines they serve
Option 2: Target databases on the basis of the disciplines they serve

Level of user support:
The BL may decide to provide user support for cross-disciplinary usage over and above that provided by the original database. The type and depth of this support may vary depending on the BL's perception of its user base.

Option 1: Provide added user support for cross-disciplinary usage
Option 2: Provide no additional user support for cross-disciplinary usage

S4. Interface

User-interface:
The interface provided by a database owner may determine whether the BL sees an opportunity to enhance the access functionality or usability of the database. If the interface is complex and specialised, the BL may find it unattractive to target the database except by providing transparent access to its existing interface. On the other hand, if the interface is too complex or is inappropriate for some identified user groups, the BL may be able to provide a simplified interface that would serve those groups.

Option 1: Rely on native database interface
Option 2: Provide a simplified interface to the database
Option 3: Provide an interface offering enhanced functionality
Option 4: Provide interfaces suitable for a wider range of users

Programmable-interfaces:
If the interface to a database supports program interaction, the BL may or may not take advantage of this capability to enhance functionality or access or to simplify the interface for new groups of users. Limited use of such capabilities might enable the creation of improved interfaces with relatively little effort; but significant use of such capabilities would require considerable design and programming expertise and resources, plus continuing maintenance and support costs.

Option 1: Do not use programmable interface capabilities
Option 2: Make limited use of programmable interface capabilities
Option 3: Take full advantage of programmable interface capabilities

S5. Interoperability

Self-describing data:
Interoperability among databases is facilitated by self-describing data, that is, by the provision of metadata to allow automated conversion and manipulation of data across databases. In the absence of such metadata, interoperability may require considerable customisation and human intervention. In order to support interoperability among databases that are not self-describing, the BL may have to develop metadata for existing data or develop automated metadata-generating processes that can supply such metadata for future data in the database.

Option 1: Rely on any existing metadata for interoperability
Option 2: Develop new metadata to support interoperability
Option 3: Develop automated processes to generate metadata to support interoperability

Semantic transparency:
Interoperability ultimately depends on the semantic alignment of interoperating databases. This is facilitated if each databases is semantically transparent, that is, if the meaning of its data is obvious or is formally described by semantic metadata. In order to support interoperability among databases that are not semantically transparent, the BL may have to develop ontologies to encode the semantics of each of the databases in question and to populate these ontologies with semantic descriptions of the data in each database, as well as obtaining or developing tools that can use these ontologies to align the semantics of these databases so that they interoperate meaningfully.

Option 1: Rely on native database semantic transparency
Option 2: Develop and populate ontologies for each database and obtain or develop tools that use these ontologies to align the semantics of these databases so that they interoperate meaningfully

Linkage-to-other-collections:
The BL may decide to link databases to other collections in order to provide interoperability in support of BL users who need to utilise multiple databases. Linkage may be at a high, dataset level or at a lower, data item level. High level linkages should be relatively easy to add to a database, but if low-level linkages are not natively provided by a database, it may take considerable effort to add them. Furthermore, if the BL merely provides transparent access to a database, it would have to add any such linkages in its own data space, virtually linking these to the data in question. Even if the BL re-hosts a database or serves as a gateway to it, the database would have to be extended to store the required linkage data.

Option 1: Rely on native database linkages
Option 2: Virtually associate BL linkage information with the database
Option 3: Re-implement the database to contain new linkages

Use of semantic standards
If a database does not utilise semantic standards, the BL may consider supplying added curatorial and technical support for such standards in a gateway or re-hosting mode, though this might require considerable effort, potentially involving the re-encoding of much of the data in the database.

Option 1: Rely on the native semantics used by the database
Option 2: Use semantic standards to re-encode and/or describe the database

Cross-domain semantic cross-walks
Depending on the anticipated cross-disciplinary and cross-domain usage of a database by the BL's users, the BL may consider providing semantic cross-walks to help its users translate data across domain boundaries and perform inter-disciplinary research.

Option 1: Rely on native semantic cross-walks (if any) provided by the database

Option 2: Provide semantic cross-walks across relevant domains

Programmable interfaces:
If a database provides programmable interfaces that can support interoperability, the BL may or may not take advantage of this capability to enhance interoperability. Limited use of such capabilities might improve interoperability somewhat with relatively little effort; but significant improvement of interoperability would require considerable design and programming expertise and resources, plus continuing maintenance and support costs.

Option 1: Do not use programmable interoperability interface
Option 2: Make limited use of programmable interoperability interface
Option 3: Take full advantage of programmable interoperability interface

S6. Ownership, funding, governance, management and contributors

Reputation
These attributes may affect the BL's interest in a given database. If the reputation of the owner of the database is good, there may be more motivation for the BL to provide access to it or serve as a gateway to it. Conversely, if the owner's reputation is poor but the database itself is considered important, the BL might be motivated to re-host it in order to improve its quality or increase its availability and prominence. In between these extremes, the BL might offer curatorial advice to the owner in order to improve the quality or reputation the database without re-hosting it.

Option 1: Improve access to a high-quality database
Option 2: Re-host an important database to improve its reputation
Option 3: Work with the owner to lend the BL's reputation
Option 4: Work with the owner to improve the database's quality

Involvement
The degree of involvement that an owner of a database has in the curation, maintenance and support of that database may affect the BL's decision as to whether or not to target the database, and if so what mode of access to provide. If an owner is uninvolved, the BL might be motivated to increase the database's visibility or timeliness; however, an uninvolved owner might not provide much support to the BL in re-hosting or serving as a gateway to the database. On the other hand, an involved owner might feel less need to have the BL support their database, though conversely, they might be more open to such collaboration.

Option 1: Work with a database's owner to improve the database's curation, quality or availability
Option 2: Interact with the owner just enough to provide transparent access

Option 3: Minimise interaction with the owner while re-hosting or serving as a gateway to the database

Accessibility:

Although they are independent attributes, accessibility of a database's owner has an impact similar to that of the owner's involvement in curating, maintaining and supporting the database. The degree of accessibility may also affect the BL's decision as to whether or not to target the database, and if so what mode of access to provide. If an owner is inaccessible, the BL may want to avoid attempting to collaborate with the owner, thereby curtailing its options for working with the database.

Option 1: Rely on owner interaction to help provide access

Option 2: Minimise interaction with the owner while providing access

Option 3: Avoid targeting the database altogether

Funding-level:

As is the case for a number of other attributes, funding-level can be seen as a challenge or an opportunity. If funding for a database is low, this poses a challenge to the BL, since the database may not maintain high quality, may be poorly supported or curated, and may have minimal resources for collaboration; however, this also creates an opportunity for the BL to add value to what may be an underfunded but important data resource. Conversely, if funding for a database is high, it may have less need for collaboration, but it may be more able to collaborate.

Option 1: Use BL funds or fund-raising capabilities to work with a database

Option 2: Use BL funds to provide enhanced gateway access to a database

Option 3: Use BL funds to re-host a database

Option 4: Partner with a database to improve or enhance its capabilities

Funding-reliability:

The reliability of funding for a database may or may not be apparent, but databases whose future funding seems more secure may be better targets for long-term collaboration, whereas those with doubtful future funding may be more suitable targets for re-hosting or at least backup. Backup strategies might include the creation of "dark archives" that are to be opened for access only if the database itself is ever shut down.

Option 1: Assume the database will continue to be supported

Option 2: Develop a backup strategy with and for the database

Option 3: Re-host the database as a hedge against its disappearance

Option A: Provide an archive for the database

Option D: Provide a "dark archive" for the database

Governance-quality:
Although they are independent attributes, the governance quality of a database has an impact similar to that of its owner's involvement. The governance quality of a database may affect the BL's decision as to whether or not to target the database, and if so what mode of access to provide. If governance quality is low, the BL might be motivated to try to improve the database's curation by re-hosting it; however, a poorly governed database might not provide much support to the BL for this. On the other hand, a database with high quality governance might have less need for BL support, though conversely, it might be more open to collaboration.

Option 1: Work with a database's governing body to improve the database's curation, quality or availability

Option 2: Interact with the governing body just enough to provide transparent access

Option 3: Minimise interaction with the governing body while re-hosting or serving as a gateway to the database

Sustainability:
Although they are independent attributes, sustainability of a database has an impact similar to that of its funding-reliability. The sustainability of a database may or may not be apparent, but databases whose sustainability seems more secure may be better targets for long-term collaboration, whereas those with doubtful futures may be more suitable targets for re-hosting or at least backup. However, a database may be unsustainable for many reasons, some of which may have nothing to do with its funding, governance or reputation: for example, it might be unsustainable because it contains ephemeral data corresponding to short-lived phenomena whose value and interest are only temporary.

Option 1: Assume the database will continue to function

Option 2: Develop a backup strategy with and for the database

Option 3: Re-host the database as a hedge against its disappearance

Option 4: Develop an archiving strategy for an ephemeral database to preserve its final contents

Option A: Provide an archive for the database

S7. Attribution & IP

Attribution completeness:
The BL may decide whether or not to target a database depending in part on whether the database maintains complete attribution and intellectual property information about its data. Regulatory and policy issues may prevent or disincline the BL from getting involved with a database whose IP information is incomplete, and doing so may compromise the BL's own reputation. On the other hand, the BL may be able to improve the IP and

attribution information in a database, offering this service as value-added to the database in question.

Option 1: Avoid targeting a database due to incomplete IP data
Option 2: Target a database (partly) to make its IP data more complete
Option 3: Target a database (partly) because it has complete IP data

Attribution accuracy:
Although they are independent attributes, accuracy of a database's IP data has an impact similar to that of the completeness of its IP data. The BL may decide whether or not to target a database depending in part on whether the database maintains accurate attribution and intellectual property information about its data. Regulatory and policy issues may prevent or disincline the BL from getting involved with a database whose IP information is inaccurate, and doing so may compromise the BL's own reputation. On the other hand, the BL may be able to improve and validate the quality of the IP and attribution information in a database, offering this service as value-added to the database in question.

Option 1: Avoid targeting a database due to inaccurate IP data
Option 2: Target a database at least in part to improve its IP data quality
Option 3: Target a database at least in part because it has accurate IP data

Attribution granularity
The granularity of IP data and attribution in a database may impact the cost and effort of maintaining, validating that information. Maintaining fine-grained attribution and IP data down to the sub-item level in a database can quickly become a management nightmare. On the other hand, distributing data with IP information that is too coarse-grained might expose the BL to legal challenge. As is the case for accuracy of IP information, however, the BL may be able to help modify the granularity of the IP and attribution information in a database, offering this service as value-added to the database in question.

Option 1: Avoid targeting a database due to inappropriate IP granularity
Option 2: Target a database at least in part to improve its IP granularity

Licensing, registration, agreements with owners
If a database has no such agreements but its intellectual property demands them, the BL may decide not to target the database, in order to avoid ethical issues and liability risks. Alternatively, the BL could take on the task of obtaining such agreements in a gateway or re-hosting mode, though this could be a significant undertaking.

Option 1: Rely on native licensing, registration, agreements with owners
Option 2: Establish appropriate licensing, registration, and agreements

End-user licensing

Similarly, the BL could take on the task of establishing end-user licensing for a database whose contents demand it but whose owner does not require such licenses.

Option 1: Rely on native end-user licensing
Option 2: Establish appropriate end-user licensing

Redaction/anomalisation of data

For databases whose contents include or demand redaction and/or anomalisation, but whose owner does not provide these facilities, the BL could decide to provide them as value-added features in a gateway or re-hosting mode.

Option 1: Rely on native redaction/anomalisation
Option 2: Provide added redaction and/or anomalisation facilities

D1. Research methodology, funding and stakeholder requirement

Required-access-granularity:
Although this attribute is independent of the supply-side access granularity attribute, its impact on the BL is similar. If the BL provides transparent access to a database, the granularity of access required by users should not be a significant factor, except in terms of the bandwidth that may be required to pass through the BL to the database itself. If the BL re-hosts a database, however, the required granularity of access will impact the kind of interface the BL must provide, as well as the bandwidth needed to supply data to users.

Option 1: Ignore granularity as being an insignificant factor
Option 2: Consider granularity when performing sizing/load estimates

Required-metadata:
Users may require many different kinds of metadata for many different purposes. In order to provide access to databases that do not have sufficient metadata for the intended user groups, the BL may have to develop such metadata for existing data or develop automated metadata-generating processes that can supply such metadata for future data in the database.

Option 1: Rely on existing metadata to support BL users
Option 2: Develop new metadata to support BL users
Option 3: Develop automated processes to generate metadata to support BL users

Required-access-to-models:
Users may require different degrees of access to the computational models that generate, validate, access, display, render, visualise and interpret data, depending on the kinds of research they are performing. If it is to provide access to databases that do not provide

sufficient access to such models for the intended user groups, the BL may have to provide such access itself; this would require considerable resources to acquire model software along with the expertise to support its use by BL users.

Option 1: Rely on the database's native access to models
Option 2: Develop in-house access to computational models for BL users

Methods:

The specific research methods employed by database users may have a significant impact on the kinds of access they will require. If the BL is to provide enhanced access to a database, serve as a gateway, or re-host the database for BL users, it may need to interact with user groups to understand which methods they employ and what these methods require in terms of database access.

Option 1: Rely on native database access, ignoring research methods
Option 2: Provide access targeted to specific research methods
Option 3: Interact with user groups to support their research methods
Option 4: Fund or conduct research to explore the relationship between BL user communities, the research methods they employ, and the requirements these methods imply for data access

Publication/distribution requirements

The BL could decide to support user needs for publication and distribution of their research results by various value-added services attached to data retrieval, such as generating automatic citations or performing translating data into appropriate forms for specific publication venues.

Option 1: Rely on native database support of publication and distribution
Option 2: Provide added support for publication and distribution

D2. Discovery methods

Search-engines

The BL might decide to provide support for specific search-engines by creating specialised metadata or other mechanisms to help such search-engines find data, whether BL re-hosts the database in question, serves as a gateway to it, or provides transparent access to it.

Option 1: Rely on native database support for generic search-engines
Option 2: Add support for specialised search-engines

Discovery metadata

Similarly, the BL could develop specialised metadata to aid discovery in specific disciplines or by specific groups of users, whether the BL re-hosts the database in question, serves as a gateway to it, or provides transparent access to it.

Option 1: Rely on native database to supply generic discovery metadata
Option 2: Add specialised metadata to aid discovery

Other discovery resources

Finally, the BL could decide to develop additional discovery resources, such as indexes and catalogues, for specific disciplines or groups of users, whether the BL re-hosts the database in question, serves as a gateway to it, or provides transparent access to it.

Option 1: Rely on native database to supply any additional discovery resources
Option 2: Create additional discovery resources for the database

D3. Query style

Expressivity:

The expressivity of the query interface required by users of a database is a fundamental design criterion for that interface. This will vary from database to database, discipline to discipline, and across different groups of users, depending on the research methods they employ. If the BL provides transparent access to a database, it must simply rely on the expressivity of the database's native query language, but if it serves as a gateway or re-hosts the database, it will need to provide appropriate query mechanisms for its anticipated users.

Option 1: Rely on the database's native query mechanisms
Option 2: Provide query mechanisms targeted to specific research methods
Option 3: Interact with user groups to support their query needs
Option 4: Fund or conduct research to explore the relationship between BL user communities, the research methods they employ, and the requirements these methods imply for querying

Desired-interface:

This is related to the expressivity attribute: The query interface required by users of a database is a fundamental design criterion. This may vary from database to database, discipline to discipline, and across different groups of users, depending on the research methods they employ. If the BL provides transparent access to a database, it must simply rely on the database's native interface, but if it serves as a gateway or re-hosts the database, it will need to provide appropriate interfaces for its anticipated users.

Option 1: Rely on the database's native interface

Option 2: Provide an interface targeted to specific research methods

Option 3: Interact with user groups to support their interface needs

Option 4: Fund or conduct research to explore the relationship between BL user communities, the research methods they employ, and the requirements these methods imply for interfaces

Required-programmable-access:
Sophisticated database users may require programmable access to a database, to enable them to use or write their own data access, data mining, and statistics gathering programs. If a database supports such programmatic interaction (for example by providing an ~API), the BL may or may not take advantage of this capability itself to enhance functionality or access or to simplify the use of the database for BL users. If a database does not provide such support but BL users require it, the BL might consider developing and providing such capability itself, though this would require considerable effort and support.

Option 1: Do not provide access to programmable access capabilities

Option 2: Provide access to any native programmable access capabilities

Option 3: Develop and make accessible new programmable access capabilities

D4. Federation

Need-to-federate:
Some users may need to federate multiple databases in order to create what is effectively a virtual database that combines the data in existing databases.

Cross-disciplinary research in particular may require the intersection of data from several disparate sources. Such federation is typically outside the purview of any single existing database, so the BL might consider supporting federation for its users, if they have a perceived need for it. Such support might include the development of specialised metadata (discussed below), the collection and provision of procedural and methodological advice on how to perform federation, or the acquisition or development of federation tools that can examine and join data schema and data dictionaries.

Option 1: Do not provide specific support for federation

Option 2: Provide methods and advice for performing federation

Option 3: Acquire or develop federation tools for BL users

Required-metadata-support:
Federation of multiple databases is facilitated by the provision of metadata to support automated mapping of data schema, joining of distinct databases, conversion and manipulation of data across databases, etc. In the absence of such metadata, federation may require considerable human intervention. In order to enable its users to federate databases

that lack such support, the BL may have to develop appropriate metadata for the databases in question.

Option 1: Rely on the database's native metadata to support federation
Option 2: Develop new metadata to support federation for BL users

D5. Cross-disciplinary usage

Cross-disciplinary-usage:
Although this attribute is independent of the supply-side Cross-discipline attribute, its impact on the BL is similar. The BL may need to maintain and provide support for users in multiple disciplines and may need to implement access mechanisms or metadata to serve users who are working in multiple fields simultaneously.

Option 1: Rely on native database cross-discipline support
Option 2: Interact with user groups to support their cross-discipline needs
Option 3: Develop cross-discipline access mechanisms for BL users

Required-metadata-support:
Cross-disciplinary research is facilitated by the provision of metadata to enable automated units conversion, semantic mapping, and other translations across discipline boundaries, thereby enhancing the native utility of a database for such interdisciplinary research. In the absence of such metadata, cross-disciplinary research may require considerable human intervention. In order to enable its users to perform such research using databases that lack such support, the BL may have to develop appropriate metadata for the databases in question.

Option 1: Rely on the database's native metadata
Option 2: Develop cross-discipline metadata for BL users

D6. Timeliness and temporal access

Required-recency:
Although this attribute is independent of the supply-side Timeliness attribute, its impact on the BL is similar. By re-hosting, serving as a gateway, or even providing transparent access to a database, the BL may introduce additional delays or latency. If a database must be accessed by some users in real-time (such as tsunami early warning data, earthquake data, or astronomical event data), the BL must consider whether its intervention in the access chain may adversely affect recency. In some cases, recency will not be an issue. In others, any expected delays will be insignificant for the expected uses of the database. If recency is expected to be important, the BL should analyse the probable impact of the BL involvement in providing access to the database and use this analysis to help decide

whether to re-host, serve as a gateway, provide transparent access, or avoid intervention in the access chain of the given database.

Option 1: Assume that recency will not be a significant factor for users
Option 2: Interact with user groups to support their recency needs
Option 3: Consider recency as part of the BL strategy for the database
Option 4: Monitor and evaluate recency when providing access

Required-timestamp-granularity

Most databases are updated asynchronously or via transactions to ensure the consistency of related data items. For many purposes, the time at which a particular data item or set of items was last updated is immaterial. However, some users may require timestamps on data items in order to know if they are timely and consistent. Users performing different research may require different granularity for such timestamp metadata, and this granularity may impact the cost and effort of maintaining the database. Unless it were to re-host all aspects of a database, including acquiring and generating its data in the first place, it would be infeasible for the BL to increase the native timestamp granularity of a database (ie, make it finer grained), since timestamp metadata must be generated at the moment a data item is entered into the database. On the other hand, the BL might reduce timestamp granularity by aggregating the timestamps for sets of data items, producing an aggregate timestamp equal to the average, earliest, latest, etc, of the timestamps for the individual items; this could be useful for some users.

Option 1: Avoid targeting a database due to inappropriate timestamp granularity
Option 2: Rely on a database's native timestamp granularity
Option 3: Perform aggregation to reduce timestamp granularity for BL users

Desired-update-method:

For some research purposes, the synchrony of the data in a database may be of particular relevance. For example, in a highly volatile database supporting real-time events that may require emergency response, it may be advantageous for individual data items representing events to be added to the database as soon as they are available, since any delay (eg, to perform batch updates) might compromise timeliness. On the other hand, if multiple data items in a database describe distinct aspects of the discipline, it may be crucial to update them all at once, by means of an "atomic" transaction, to ensure that they are consistent. These user requirements for a database impact its data submission and update strategy, which are the responsibility of its administrators. Unless it were to re-host these administrative tasks, it would be infeasible for the BL to change the update method of a database.

Option 1: Avoid targeting a database due to inappropriate update method
Option 2: Rely on a database's native update method

Required-temporal access:

Although this attribute is independent of the supply-side Temporal-depth attribute, its impact on the BL is similar. Access to previous values in a database may take several forms, including frozen, historical information, a "snapshot" of current values, multiple versions of data (ie, previous snapshots), access to specific multiple temporal epochs, or the ability to reconstruct data at any desired time in the past. If the BL re-hosts or serves as a gateway with enhanced functionality, it may be called upon to provide temporal access beyond that which is available directly from the database itself. This could require check-pointing previous versions of the database, maintaining transaction histories, and enhancing query capabilities to specify temporality.

Option 1: Rely on the database's own temporal storage and access
Option 2: Serve as a gateway offering enhanced temporal queries for BL users
Option 3: Host temporal storage and appropriate access mechanisms for BL users