ECP 2005 CULT 038097/Bernstein

Bernstein

No 4, ref. D1.1, Integration blueprint

Deliverable number	*D1.1*
Dissemination level	*Public*
Delivery date	*14 February 20007*
Status	*Final*
Author(s)	*The Bernstein Consortium*

e*Content*plus

This project is funded under the *e*Content*plus* programme[1],
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

[1] OJ L 79, 24.3.2005, p. 1.

The Bernstein Project

The Integration Blueprint

Contents

[1] Introduction.

This paper examines integration and knowledge management models and proposes solutions for an integrated workspace for allowing scholarly access to watermark databases.
The paper includes:

- A list of fundamental functions and core features of the integrated workspace user interface
- Evaluation of the efficiency of storage and data models
- Comparisons between model types

[1.1] Background

The objective of the Bernstein project is to create a Europe-wide integrated digital environment for the expertise and history of paper. The project will interlink all existing European databases of paper reproductions, make their content accessible to specialised image processing tools for the measurement of paper features, and provide an interface to the digital resources of domains related to paper studies or by which the knowledge about papers can be enriched and contextualised.

Interoperability between resources would substantially increase their usage and impact, as compared to their independent existence. In order to avoid major modifications of existing resources and ensure ease of scalability, integration will not rely upon forcing each database and tool to conform to a single standard, but will be provided by a versatile interface able to support the individuality of all resources. The user will be able to interact with a multitude of contents and content manipulators through a single integrated workspace.

The Bernstein project has the ambition to generate the conceptual and technical infrastructure to enable the access to paper expertise to the broadest range of users, accommodating multiple usage scenarios. The principal targeted user communities are the historians and the cultural heritage conservators, although other areas where paper identification capabilities are required are expected to benefit from the project, such as the art market, forensic science, security research and the paper making industry. Therefore the focus is on historical paper, while applications to modern papers are not excluded.

Work package 1 will create an “Integrated workspace” which allows access to watermark descriptions and images. It is fundamental to the Bernstein project and will provide the digital environment necessary for the integration of the current distributed databases. Specifically, this is an Internet application that will allow communication between components of the system, will harmonise data formats and provide a unique user interface for accessing data. The integrated workspace will provide the necessary infrastructure for integrating various content repositories and content processing tools. This paper examines integration and knowledge management models and proposes a solution for the Bernstein project.

At present there are four on-line databases which cover the major digital resources on paper studies available. Altogether the four databases contain 120.500 digitised watermarks (metadata and images) which are distributed as follows: LABW, Germany (95.000): KB, Netherlands (16.000): OEAW, Austria (8.000) and NIKI, Italy (1.500). These collections represent the reference material for historical studies on paper and watermark expertise.

Within the databases there is additional metadata providing information about the measured characteristics of the paper and classification of the watermarks. There is also information

about the date and place of production of the reproduced papers and about the documents for which they are used.

[1.2] Information on present databases

The following describes the information which is available within the present databases. The four existing on-line databases are:

- "Piccard" (provided by LABW, Germany): http://www.landesarchiv-bw.de/piccard/start.php
- "WILC" (provided by KB, Netherlands): http://watermark.kb.nl
- "WZMA" (provided by OEAW, Austria): http://www.ksbm.oeaw.ac.at/wz/wzma2.htm
- "NIKI" (provided by NIKI, Italy) http://www.wm-portal.net/niki/index.php

The following fields are common to all four databases: 'reference number', 'image reference', 'distance between chain lines', 'remarks', 'height in mm', 'date',' place of origin' and 'repository'.

Regarding 'date', there is two fields (begin and end) in Piccard and WZMA and only one field in the other databases.

Regarding the 'place of origin' and 'repository', this information is stored in one field in WZMA and in two fields in the other databases.

The 'width in mm' of the watermark cannot be found within WILC and the classification is not stored in NIKI. The 'position', 'between the chain lines', and the 'twins' are not stored in Piccard and WZMA.

Regarding the motif groups, the name of the motif group is stored in different languages (English, German and French) in Piccard. The signature of the paper does not appear in NIKI.

Table 1 shows a summary of existing data which is available in at least 2 of the current databases. Each database also contains additional information which is not expressed in the table, for example, WILC also include information on paper size.

	Piccard	*WILC*	*WZMA*	*NIKI*
Watermark				
Reference number	X	X	X	X
Image reference	X	X	X	X
Distance between chain lines	X	X	X	X
Remarks	X	X	X	X
Height in mm	X	X	X	X
Width in mm	X		X	X
Classification (motif group)	X	X	X	
Position		X		X
Between the chain lines		X		X
Twins		X		X
WZMA_Nr	X		X	
Motif group				
Class hierarchy	X	X	X	
Name in English	X	X		
Name in German	X		X	
Paper				
Date (if available)	X	X	X	X
Place of origin	X	X	X	X
Repository	X	X	X	X
Signature	X	X	X	

Table 1 - Information provided by the four databases (only fields common to more than one database have been included)

[2] Fundamental Functions and Required Features of the System

[2.1]Use cases

It was decided to formulate the fundamental functions of the system by defining an all encompassing set of situations faced by users or 'use cases'. These use cases are intended to define the ways a user would want to interact with the system, the situations within which they would wish to do this and how they would like the results presented. The use cases were requested from all partners within the project and also from a selection of other interested parties as defined on the 'twiki' (http://www.bernstein.oeaw.ac.at/twiki/bin/view/Main/TWikiGroups).

The use cases have been split into the following sections:

- Integration - use cases related to data sources and methods of integration of these sources.
- Search - types of search that the system will be required to support
- Display - display mechanism and formats required
- Statistics - various methods for processing the results of a search to provide clarity or further information

Most of the use cases submitted are included in this integration blue print document. Those that could not be supported were removed. Those that were removed were: searches for medieval symbolism and searches for watermark counter-marks: this information is not available in the current databases and provision to add this data is not covered in any of the Bernstein work packages.

In addition, it is important to stress that some of the included use cases will only be supported if the underlying data is available in the original databases, for example, in use case 20, searching for water mark twins is only possible if the twin information has been recorded in a field in the database.

This project is intended to integrate each of the watermark databases, to link to other related databases, to provide a suitable mapping between the same terms in different databases independent of language and to integrate useful tools to provide expertise.

To fulfill all aspects of all use cases, the watermarks would need to be entirely re-catalogued; this is outside the scope of the project. In recognition of this need, as highlighted in the use cases, a new method of watermark description will be developed - the 'component model'. This is discussed further in section [3.2].

[2.1]Use Case List

The Integration should:

1 Provide results at a reasonable speed
2 Be accessible to the world outside Bernstein via a machine2machine link
3 Perform image scaling
4 Have appropriate error treatment
5 Have a ISTC link
6 Have a Briquet Search
7 Use a Bernstein Code for each watermark

The Search should:
8 Provide Multilingual access
9 Be reliable
10 Provide combined searches with logical operators
11 Avoid ambiguity due to subjective classification
12 Account for synonymy – database side
13 Account for synonymy – user side shape ambiguity
14 Search for specific elements within the databases
15 Search for elements within a watermark
16 Inform users if they are searching for watermarks not in the databases
17 Allow for the heterogeneous content provided by the different databases and draw the user's attention to this
18 Use terminology that is understood by users
19 Be able to combine other factors such as measurements to narrow down search results
20 Be able to search for watermark twins
21 Be able to search for bibliographical references

Display should:
22 Incorporate all functions into one workspace
23 Provide multilingual interfaces
24 Provide a method for a user to store results

The statistics that should be provided are:
25 General statistics on the results of a search
26 The facility to export results
27 To provide information on the dating of paper
28 To provide indications of document authenticity
29 Plotting results sets on to maps
30 Plotting bibliographic references on a maps

Off Line Tools:
31 Image Processing

[2.3] Search / Display

At present there is three databases which provide textual search (WILC, Piccard, NIKI) and three which provide textual hierarchy/ image based browsing (WZMA, WILC, Piccard). It is felt that successful implementation of this project will need to offer both of these facilities. It has not, as yet, been discussed and decided as to which of the fields in the present databases will be made available to search from within the Bernstein workspace.

It has been decided that the architecture would need to support three types of searches:

- **Text.** Searches of varying complexities, combinations of visual and quantifiable parameters combined via boolean operators the user will be able to perform complex textual searches which can be linked by logical operators – use case 10.
- **Constructed image.** It will be possible to create a representation of the query watermark for use in finding similar or exact matches in the target databases.

- **Combined.** Using text to construct an image. Text based queries will either be mapped silently into this query description as above, or used as input to automatically construct an image in the graphical representation, which can be viewed by the user before being submitted as a search.

From the use case it is important that the searches provide multilingual access as discussed in [2.5]. Many of the other use cases are self explanatory, for example, 10 which provides combined searches with logical operators' and 22 which incorporates all functions in one work space and so on; more detail is provided below for more complex use cases.

Use cases 11, 12 and 13, all discuss the problems of ambiguity of shape and synonyms, for example, a user may be searching for a shape which he thinks is a ribbon. In one of the databases it might be classed as a rope (similar shape) whereas in another it might be classed as a snake. In the Bernstein workspace there should be a way to map these synonyms and remove the ambiguity. The terminology used in the Bernstein Project will be at a suitable level so that it is of use to watermark experts and still understandable to novel users as highlighted by use case 18.

In an additional search, the user may want to look at all watermarks that contain this ribbon / snake / rope element but this could be more likely to occur within a watermark that contains more prominent elements such as a snake with a dagger or a rope with an anchor. In the present databases these watermarks could be classified under the motifs: anchor or dagger, not ribbon / snake / rope. Therefore, it is impossible to search directly for the ribbon or snake or rope. This issue was highlighted in use case 14 – a desire to search by elements. This issue is reconfirmed in use case 15 – if a user has an incomplete watermark, a common occurrence for the art historian, he may only have a section of the watermark with a snake on it but the dagger is missing. The watermark may be classed as a dagger watermark; therefore, he has no way to find it as he does not know that the dagger exists.

In use case 16 the user wishes to know if his watermark does not exist in the database, for example if a user searches for all watermarks that contain both a bull's head and an anchor, he wants to be told that there is no record of this watermark rather than be given all bulls head and all anchor watermarks. This relates to the boolean operator as discussed in use case 10.

Use case 21 describes two methods for accessing bibliographical data. In the first the user wishes to input a search term within one of a number of defined fields, for example, author, title, publisher, subject, classification, he expects to receive data relating to paper history and watermarks. In the second case a user already has a list of watermarks from a search and is interested in bibliographical context relating to these watermark locations or motifs. The results for both of these methods would be displayed in a sortable list.

The user will wish to store their results – use case 24. This use case means that the Bernstein workspace will support the ability for the user to log in to the workspace where they will locate their own area which will contain a record of information they have stored from past visits.

[2.4] Statistics

The use cases in the statistic section revolve primarily around post-processing of search results. For each statistical package, information is derived from a search and fed into appropriate statistical processes and data added, where appropriate, from other sources such as the bibliographic databases. The results will be displayed within the Bernstein workspace.

There will be a statistical package that will analyse the data in the Bernstein databases and produce statistics relating to: population, data distribution, dates, number of motifs and percentage catalogued within the component model. These statistics will also be able to be performed on datasets produced by users, use case 25. In addition, the user will be able to determine the percentage, of the whole dataset, within which his search falls.

The tools for statistical analysis, such as paper authenticity and mapping of data, will be developed as part of the Bernstein work packages and will be integrated into the architecture as they become available. Links will be made with further bibliographical databases, and searches will be performed on those databases as dictated by the search results.

[2.5] Multilingual User Interfaces

Language support will be provided within Bernstein products in respect of:

	English	French	German	Italian	Russian	Spanish
Databases						
- user interface	X	X	X	X	X	X
- data display	X	X	X			
- search	X	X	X			
Expertise tools						
- user interface	X	X	X	X	X	X
- data display	X	X	X			
Cartography						
- user interface	X	X	X	X	X	X
- data display (*)	X	X	X			
- search (*)	X	X	X			
Bibliography						
- user interface	X	X	X	X	X	X
- data display	X		X			
- search	X		X			
Dissemination						
- kit software						
--- user interface	X	X	X	X	X	X
--- data display	X	X	X			
- kit documentation	X	X	X	X	X	X
- handbook	X		X			
- advertising	X					
- technical doc.	X					

Table 2 - Language support within the Bernstein project

(*) Depending on the availability of georeferences

The system will provide multilingual explanations and help pages in six different languages: English, French, German, Italian, Russian, and Spanish. Querying of the data and results given

will be supported in German, French and English. This means that watermark descriptions will be mapped to each of these three languages (see WKP2).

[3] Integration Models

[3.1] Data storage

[3.1.1] Evaluation of Models

The data from the current databases will be integrated on the basis of one of the following models: a distributed approach, a centralised approach or a combination of the two. These are discussed below:

Distributed Approach

The databases are accessed in their existing forms from their current locations. When a user submits a query to the central Bernstein interface (integrated workspace), each database is searched via the internet in real-time. This is achieved using a machine-to-machine protocol (e.g. SRU, Z39.50, OpenSearch). Results from the individual databases are assembled, sorted and formatted for display. These will be presented to the user by the central Bernstein service.

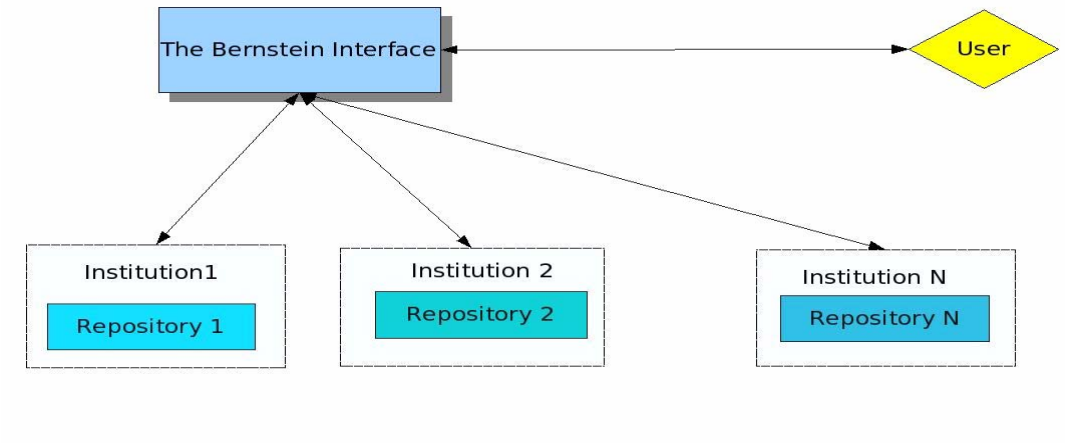

Diagram 1 – The Distributed Approach

Centralised Approach

In this model the data is exported from the existing databases and ingested into a single 'Super' database. Ideally, this database is situated within close geographic proximity to the server providing the central Bernstein service for reasons of efficiency. The export and ingest processes will take place before the service is launched and the data can be regularly updated using this

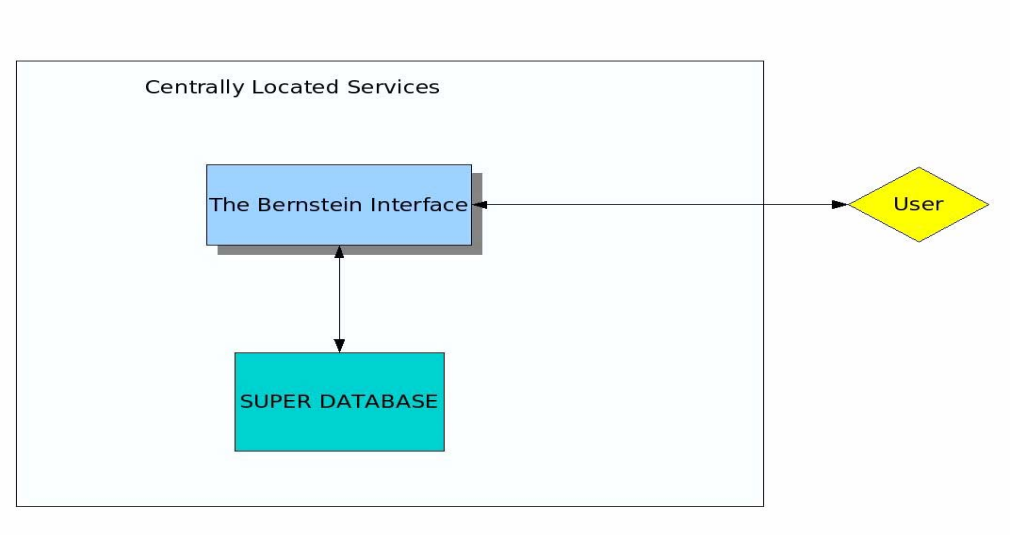

Diagram 2 – The Centralised Approach

same method. When a user submits a query to the central Bernstein interface, the single 'Super' database is searched and the results formatted and presented to the user.

Combined Approach

The aspect which differentiates the combined approach from the previous models is that while the databases are kept separate from each other (as in the distributed model), it is also possible to host copies co-located with the server responsible for the central Bernstein interface. Some databases could be held centrally and others held by their owning institutions. Regardless of the location, each database would be accessed by the same interoperability protocols, allowing the migration of the database between remote and central hosting as desired. When a user submits a request to the central portal, each database is queried individually and the results compiled, formatted and presented to the user.

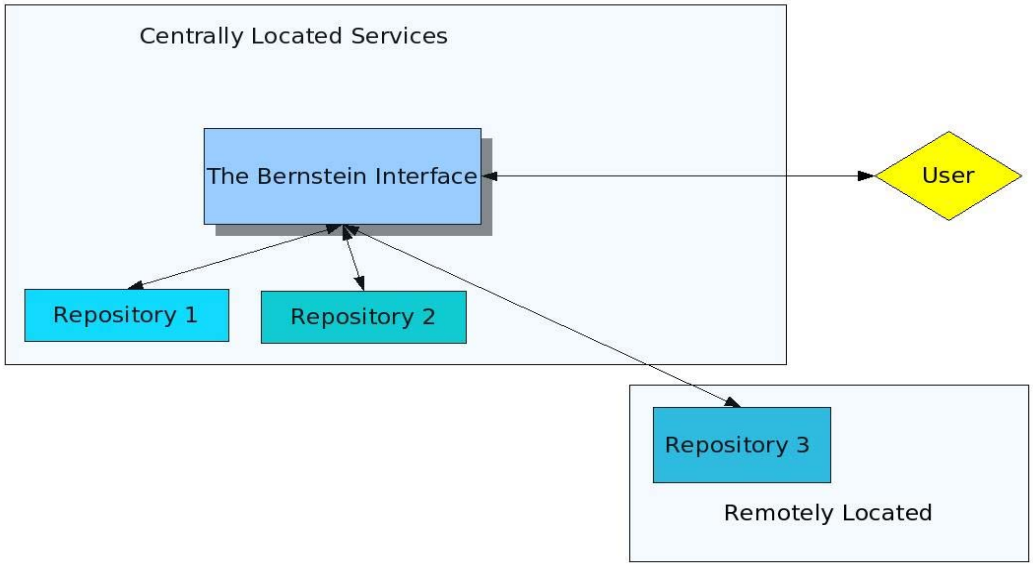

Diagram 3 – The Combined Approach

[3.1.2] Comparison of the Models

The first solution, the fully distributed approach, gives the control and responsibility of the databases to the institutions that created them. From a sustainability point of view this is a positive aspect, as the institutions are more likely to persist beyond the time frame of the Bernstein project. However, it does mean that each institution is responsible for ensuring that the information is constantly available which may require additional technical expertise. For example, any changes to the database, such as change of database structure or changes of database type, must not disrupt the interaction with the Bernstein portal.

Institutional servers are normally protected by some kind of firewall; each institution would need to enable ongoing access across this firewall. It must be possible to query each database from the central server and it may not be possible for the institution to provide this access for either technical or administrative reasons. This may be seen as a barrier to entry for new databases wishing to join the consortium, and therefore it could be a limiting factor for the extensibility of the architecture. This solution also means that potentially large data sets would need to be transmitted from the individual databases and processed on the central server each time a search is performed – this would extend the time taken for each search and goes directly against the needs stated in use case 1.

In the second solution, the centralised approach, all data will be stored in one central database. This approach means that no machine-to-machine access is needed and resulting datasets for processing would not need to be transmitted across the internet. This would result in a much shorter search time, as desired in use case 1. It also removes the need to deal with access across firewalls and access via machine-to-machine protocols. This solution means that the type of central database used to store the information can be chosen to best meet the needs of the Bernstein system and would not be dependent on legacy choices of the original database owners.

Changes to the original databases or data can be made as required and will not affect the exported data. This means that changes to the original data will not be apparent from within Bernstein until the data is re-exported – there would need to be some automated method for this update process. This solution imposes serious sustainability issues upon the Bernstein host. There is a commitment from the Bernstein project to ensure access to this information for five years after the project, but not indefinitely, after which time the information could become unavailable. There are also issues surrounding copyright and intellectual property rights concerning the watermark descriptions – the institutions may not be allowed to give away copies of their data to other third parties for legal reasons.

The combined approach has many of the positive aspects of the previous two models. It will provide extensibility and scalability of the infrastructure, enabling future prospective partners to join either as a remotely searched database conforming to the interoperability protocol, or as a centrally hosted database. For some institutions, it is essential that they be able to host their own data due to intellectual property requirements. However, for institutions without significant resources or expertise in technical matters, it is equally important that their data be included. The distributed databases will always be up to date – any change to the database will instantly be reflected in the main Bernstein interface. Centrally hosted databases can be updated as deemed appropriate by the institution. The combined model features the best sustainability approach. Institutions able to provide an interface to their own data will be more likely to maintain it indefinitely, and Bernstein will provide access for at least 5 years after the end of the project for any other institution or individual with any other database. The

combined approach allows for third party access to the data where permitted as even the centrally hosted databases are accessed in the same way as the remote ones – for example the TEL project may wish to incorporate the watermark databases into their portal. Finally, it is possible for databases to switch between being centrally and remotely hosted, as priorities may change over time.

Where possible, the databases should be accessed from their present locations by machine-to-machine protocols. If preferred by the owning institution, access to the information can be provided at a central server on a copy of the data exported from the original database.

In addition to the models above, the databases are to be integrated by providing whenever possible the references to similar watermarks in printed media repertories such as those of Piccard and Briquet. Bernstein will thus offer back-ward compatibility to the standard works for paper studies in non-digital media. Further to these references the main classes of IPH could be used in different databases to offer other possibilities of integration and integrated searches.

Preferred Bernstein Architecture

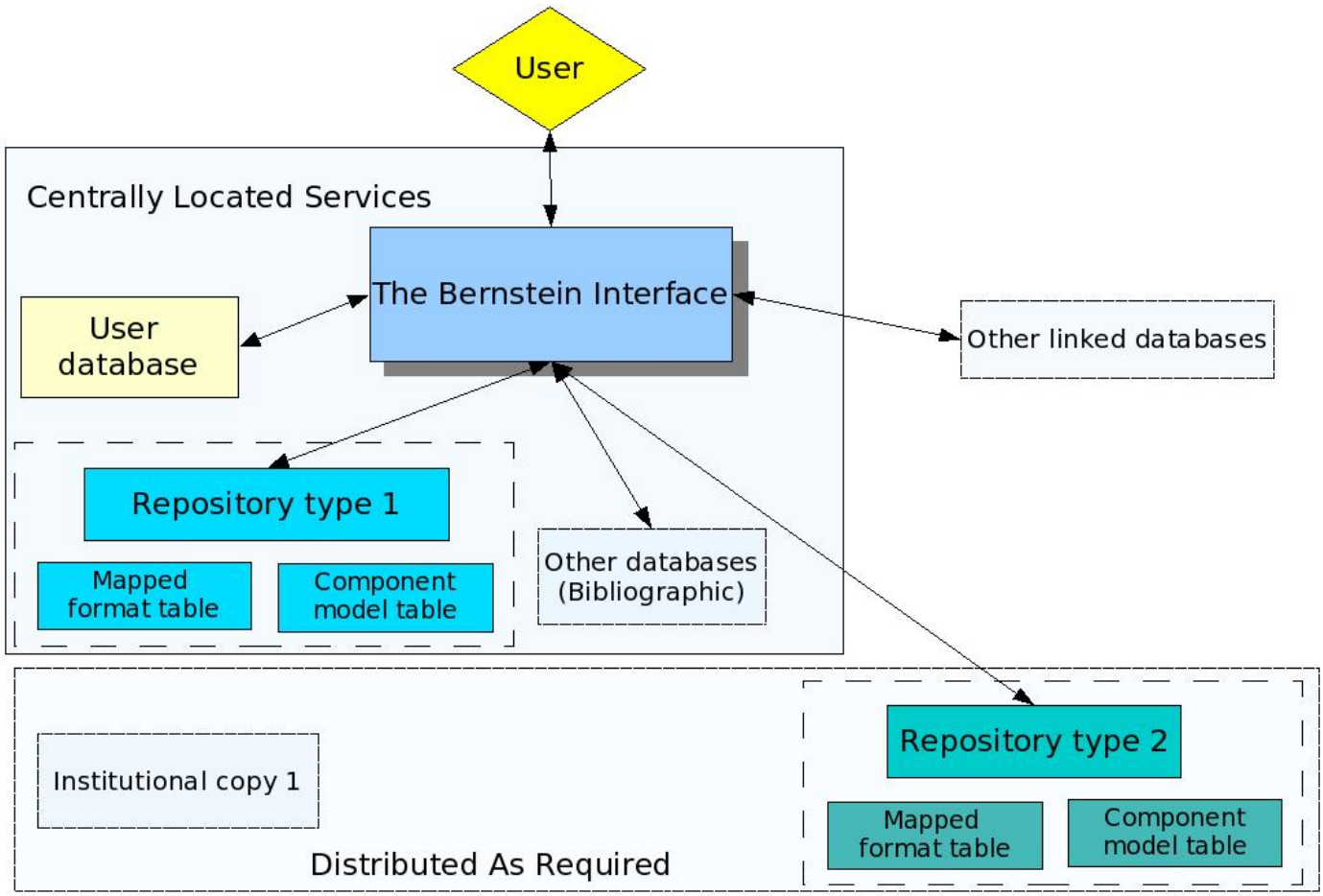

Diagram 4 – The Bernstein Approach

[3.2] Data Models

It is also important to consider the semantic differences between watermark descriptions coming from different databases, as opposed to the protocol level interoperability previously considered. Each database has been developed independently, using different hierarchies for textual description or classification and different measurement standards and methodologies for numerically oriented fields. In order to successfully retrieve all of the appropriate records, some level of data harmonisation must be considered.

[3.2.1] Evaluation of data models for watermark description

There are several ways the data can be harmonised:

Status Quo

No additional descriptive work would be done and mapping work would not be performed between the current descriptions in the databases. The databases would be integrated as they stand.

Coordinate the data via mapping

The data contained within the original databases is organised into a specific hierarchy. Those hierarchies are specific to each database. To coordinate the databases for integration, each of the hierarchies will be mapped against the other hierarchies.
The classes will be mapped to the equivalent classes in other databases. For example, rope to ribbon to snake. Mappings can also be made via identifiers such as Piccard or Briquet numbers.

Due to the fact that the different databases have classified the watermarks in different ways, direct mapping may not always be possible. Watermark experts will decide the most suitable matches.

The database will also be mapped to the three different search languages in order to fulfill the multilingual aspect of the search (use case 8). Mappings will be created to deal will the aspects of ambiguity (use case 13) and synonyms (use case 12).

The Component model

A watermark can be described in terms of one or more distinct objects (henceforth 'components') and their spatial inter-relationships; for example, a cross above a bull's head both being enclosed within a circle. That would be a different methodology to the existing mechanisms for description which rely on the hierarchy to capture those relationships. By capturing the watermark descriptions in this manner, access to the components can be implemented via different semantic hierarchies as appropriate to the user in terms of skill, background and language.

Each component will have its own description as appropriate to the type of component and the individual watermark. This description will be made up of various attributes; for example, a bull's head may have attributes of eyes, horns and nostrils. Each watermark would thus be constructed from one or more components plus any relationships such as 'above', 'left-of', or 'within'. The list of components, attributes and relationships will be decided by watermark

experts. The watermarks will then be described by catalogers in terms of this model. Within the catalogue, the different types of components will be referenced by a number to ensure consistency to enable synonym searches and ensuring language independence.

[3.2.2] Comparisons of the Models

Given the inherent dependency of integration on at least some degree of content harmonization, it is impossible to fully integrate these databases without some degree of mapping work. The databases were originally developed independently of each other and, therefore, have been constructed in entirely different ways; the databases may be semantically similar but they are syntactically very different, using very different structures to record their data and different terms to describe the same thing. It is not possible to integrate the databases as they are - thus removing the status quo option.

The mapping option will deal with issues of multilingualism and the integration of the different cataloguing hierarchies to a certain extent, however, there are likely to be some irresolvable differences in existing descriptions as the existing databases have been developed completely independently. Whilst this task is complex it does not require the full re-description of any watermarks and hence requires the least amount of data input.

Using a mapping layer between the databases will provide some resolution for the use of different terms, however, there are several situations where it will not be sufficient to cover all of the projected use cases. Searches for incomplete descriptions will be limited, and the correctness of any search will be impossible to guarantee as it may not be possible to create the mapping in enough detail to capture all of the different descriptive semantics. Even mapping each individual record, rather than the classification within the hierarchy, would not be sufficient to fulfill many of the use cases provided.

Another option would be to create a new system of description, the component model, thereby creating a descriptive standard for watermarks – not a trivial task in itself. Then re-describing each watermark would be required – this significant amount of work was not foreseen at the outset of the project.

A compromise solution is needed. The mapping work will be performed between the databases to reduce the ambiguity of terms as described in use cases 9 and 11. Mappings will be created to deal with the synonyms as highlighted in use cases 12 and 13. It will provide users access to search all the watermark databases within one workspace. For each of the databases, certain watermarks will also be redescribed using the component model; for example, all bull's head motifs could be redescribed using the component technique. This will mean that users will be able to search for partial watermarks, use case 15, and also specific elements within watermarks. It will completely remove the issues of synonymity and multilingualism as the data will be stored as numerical codes rather than descriptively – completely satisfying use case 8, 9, 11, 12 and 13. When searching, both descriptions, component and non-component, will be accessed from each of the databases. The results will be presented to the Bernstein workspace in separate lists. As time and resources allow, more watermarks could be redescribed using the component model if this model proves successful.

[4]Conclusions

The fundamental functions of the system have been defined by an all encompassing set of use cases.

Use cases relate to the integration of the systems component databases, the types of search the system will be required to support, the display mechanisms and formats required, and various statistical methods for processing the results to provide further information.

The databases could be integrated via a distributed approach, which would link the current databases to each other or by a centralised approach that would export all of the data into one central database. A combination of these two approaches will be used in the Bernstein project. Data will either be exported from the original databases and stored in separate centalised databases on a system local to the Bernstein interface or the Bernstein interface will link to the databases at their host institutions. That interface will also link the tools required for statistical analysis and link to any external databases required such as the ISTC and the bibliographic databases.

The information present on the current databases will be mapped to each other, this will ensure that a Bernstein user will be able to search for all watermarks currently catalogued in various databases within one workspace. A subsection of watermarks will be reclassified in terms of the component objects. This additional description will ensure that all user situations that have been defined will be supported.