RAL-TR-2005-003

CRISs, Architectures and CERIF

K G Jeffery

18 January 2004

Council for the Central Laboratory of the Research Councils

© Council for the Central Laboratory of the Research Councils

Enquiries about copyright, reproduction and requests for additional copies of this report should be addressed to:

Library and Information Services
CCLRC Rutherford Appleton Laboratory
Chilton Didcot
Oxfordshire OX11 0QX
UK
Tel: +44 (0)1235 445384
Fax: +44 (0)1235 446403
Email: library@rl.ac.uk

CCLRC reports are available online at:
http://www.clrc.ac.uk/Activity/ACTIVITY=Publications;SECTION=225;

ISSN 1358-6254

Neither the Council nor the Laboratory accept any responsibility for loss or damage arising from the use of information contained in any of their reports or in any communication about their tests or investigations.

C L R C

CERIFTG/ 01 /04

RUTHERFORD APPLETON LABORATORY AND DARESBURY LABORATORY
BUSINESS AND INFORMATION TECHNOLOGY DEPARTMENT

EUROCRIS

CRISs, Architectures and CERIF

Keith G Jeffery

18 January 2004

1. INTRODUCTION

CRISs exist; some of them have existed for a long time. They have different purposes / requirements, data models and implementation platforms. However, there is a general recognition that it would be useful to provide a homogeneous view over these heterogeneous CRISs, especially in the ERA (European Research Area) context. There are several ways to implement a solution to the problem of homogeneous access to heterogeneous CRISs. This short paper attempts to present and compare them. Some background information is attached as an Annex; this provides explanations and definitions of terms and also some previous experiences.

Apart from harvesting, which ignores the syntax (structure) and semantics (meaning) of the data and just does text string searching, all architectures rely on a predicate query over a known schema (available or derived by schema reconciliation) thus allowing search terms or values to be related to an entity/attribute and thus to a domain. Example: the string 'green' could occur under attribute 'family name' in entity 'person' or within attribute 'abstract' or 'title' in entity 'project'. The use of query under a schema ensures that the query is meaningful and should have adequate recall (coverage) and relevance (precision).

Most techniques rely on navigational metadata to access hosts of CRISs. The catalog techniques use in addition associative descriptive metadata to perform the first pass search – rather like the harvesting technique, but using structured and meaningful data under entity/attribute sanction. Those techniques with server-side or client-side wrappers require schema metadata to perform schema reconciliation. Although CRISs of these architectures usually do not use CERIF, CERIF can be used with advantage to define the database schemas.

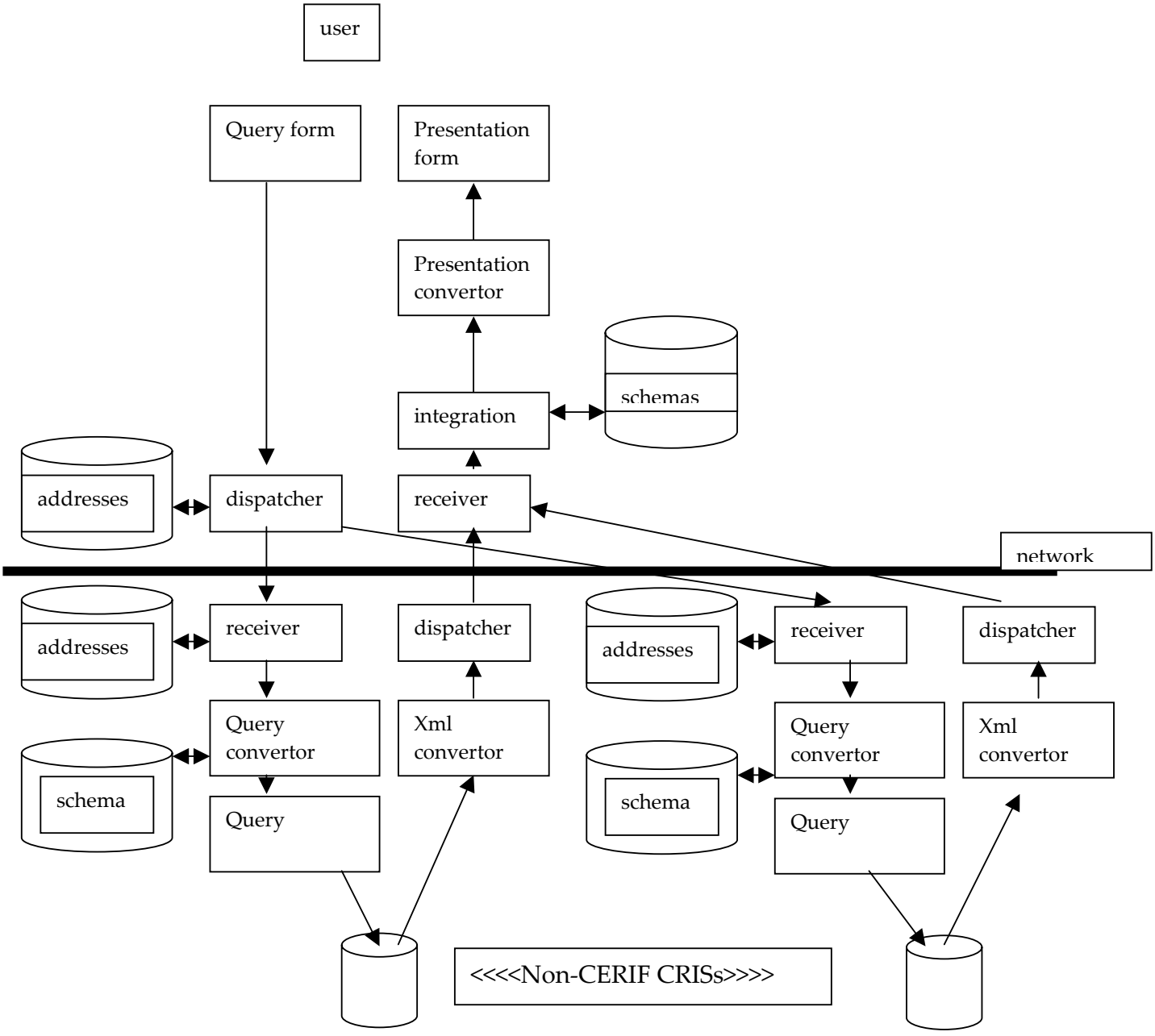
One technique (Full CERIF) has a uniform assumed schema and so has no need for metadata nor schema reconciliation. However, it relies on each host either being full CERIF compliant or providing (and maintaining) a full CERIF version of the host database.

CERIF provides the optimal solution in the full implementation. All other techniques benefit from using CERIF to define schemas or export formats.

This paper describes the different architectures. Annexes provide references to (1) Metadata, (2) CERIF, (3) the EXIRPTS Protocol and in (4) the complete text (since the report is not easy to obtain) of the ERGO Group Report to the EC recommending the ERGO 2++ architecture.

2. REMOTE WRAPPER

2.1 Architecture



2.2 Description

This architecture provides a simple user query interface to multiple host CRISs. Each host CRIS has to

(a) provide address to the client dispatcher database

(b) provide software for query conversion to local host DML (using host schema)

The use of XML to encode answers (an addition to the basic architecture but indicated in the diagram) dispatched provides some syntactical uniformity but no uniformity of character set, language, semantics. Uniformity in these other aspects can only be achieved through a canonical data model (CERIF). Unfortunately XML cannot represent the full semantics of CERIF, because it represents hierarchies and CERIF represents a directed graph.

2.3 Metadata

This architecture uses schema metadata for query conversion and answer integration. It uses navigational metadata for access to hosts. This architecture does not use CERIF metadata.

2.4 Process

The user inputs a query through a supplied web browser form of the kind 'find the string "widget" anywhere in any host database'. The dispatcher sends this query in some protocol [email | ftp | message] to all hosts with address entries in the address database. Each host converts to its own DML (using the host schema) and executes the query. The results at each host are converted to XML (an addition to the basic architecture but indicated in the diagram) and dispatched back to the user who receives one XML file per host (each with differing syntax and semantics). The integrator takes the result sets and using the host schemas (or preferably XML DTD equivalents of the schemas) reconciles them to a uniform result set which is converted for end-user viewing by the presentation converter.

Advantages and Disadvantages

(a) user needs only web browser and simple query form

(b) host has to write query converter

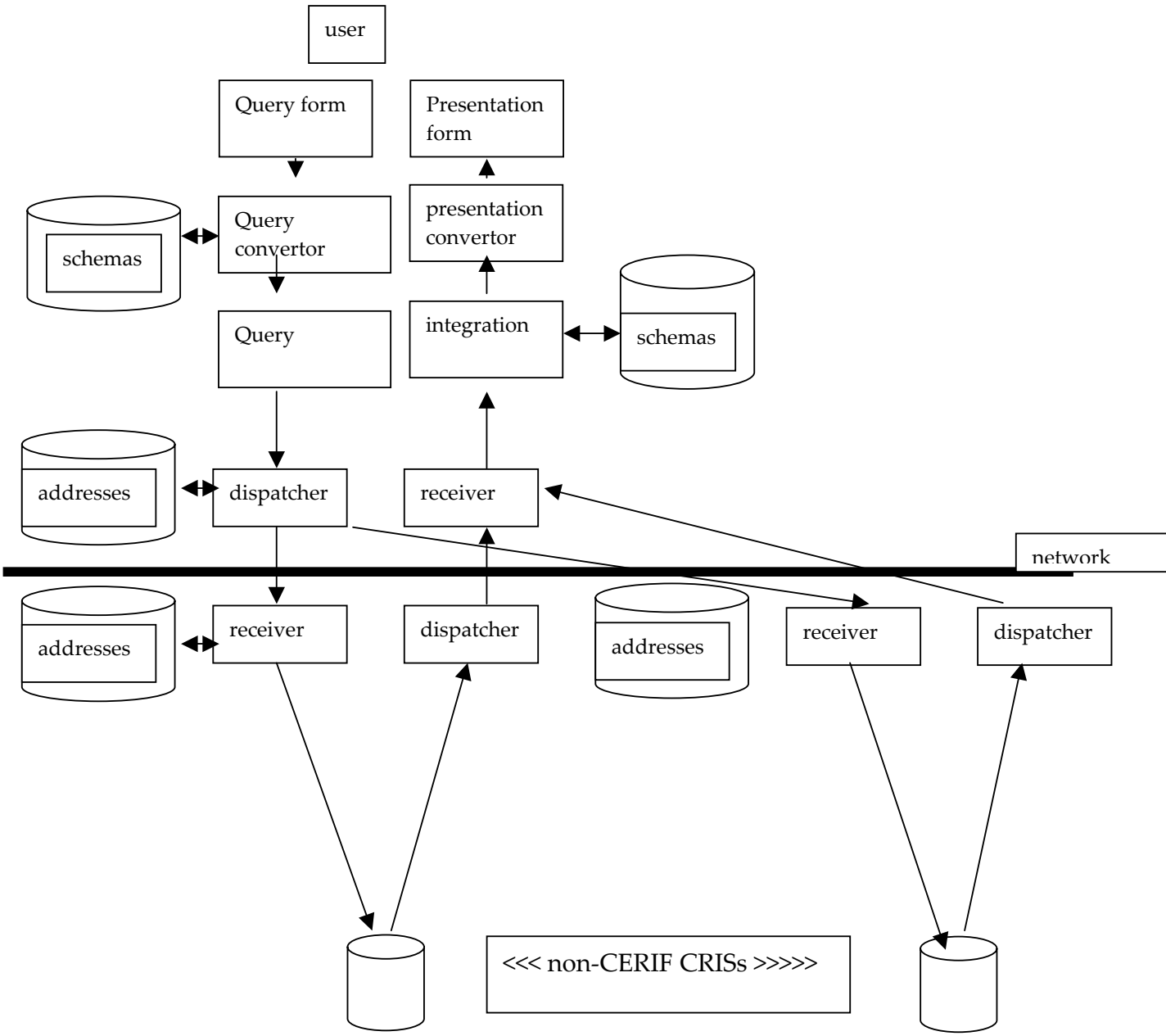
(c) host has to write XML converter (to a specific XML DTD?)

(d) query expressivity very limited

(e) user client has to write an integrator for the answers

3. LOCAL WRAPPER

3.1 Architecture



3.2 Description

In this architecture the hosts have only to provide a receiver and dispatcher; they receive queries in their own DML and dispatch results in their own data model. All conversion responsibility is on the client. The client provides queries for each host from the user query, mediated by the host schemas and integrates the results from each host, using their schemas, to an answer for the end-user presented through a user-defined presentation converter (e.g. XML, HTML….).

3.3 Metadata

This architecture uses schema metadata for query conversion and answer integration. It uses navigational metadata for access to hosts. This architecture does not use CERIF metadata.

3.4 Process

The end-user generates a query in some arbitrary language, using a query refinement interface and web form. The client software converts the query to the target DML for each host using the host schemas stored (and updated by the hosts) at the client and dispatches them using the addresses database. Each host receives a query in its own DML, executes it and returns the result in its own form via the dispatcher to the client receiver. The integrator takes the result sets and using the host schemas reconciles them to a uniform result set which is converted for end-user viewing by the presentation converter. CERIF could, with advantage, be used as the uniform schema for result integration.

3.5 Advantages and Disadvantages

(a) each host has only to supply and update its schema to the client (all clients if there is not a central query server)

(b) each host has no software to provide except receiver and dispatcher

(c) the client (if it is a central service) has a very large workload

(d) if there is no central service then each client has to have all schemas supplied and updated

(e) the client software has to include a complex query refiner

(f) the client software has to include multiple complex query converters

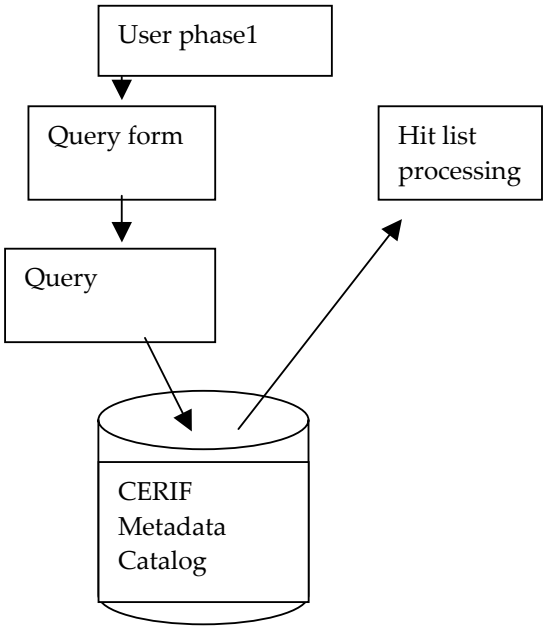
(g) the client software has to include a complex answer integrator

(h) the client software has to include a presentation converter (complexity depends on specification of presentation required and complexity of the answer structure)

4. CATALOG

4.1 Catalog Only (ERGO Pilot)

4.1.1 Architecture



4.1.2 Description

This architecture provides a canonical subset data model – CERIF metadata model – with one character set, one language, one syntax (structure) and one semantics. This provides the homogeneity.

4.1.3 Metadata

This architecture uses associative descriptive metadata (CERIF metadata catalog)

4.1.4 Process

The CERIF metadata catalogue is populated from all hosts which provide a local converter from their data model to CERIF metadata (one character set, one language, one syntax (structure) and semantics (meaning)). The end-user has a query form which queries the catalog and obtains a 'hit list' of results. Experience indicates such results satisfy ~ 80% of queries; however if more detail is required the architecture provides the capability for accessing the hosts (see next section).

4.1.5 Advantages and Disadvantages

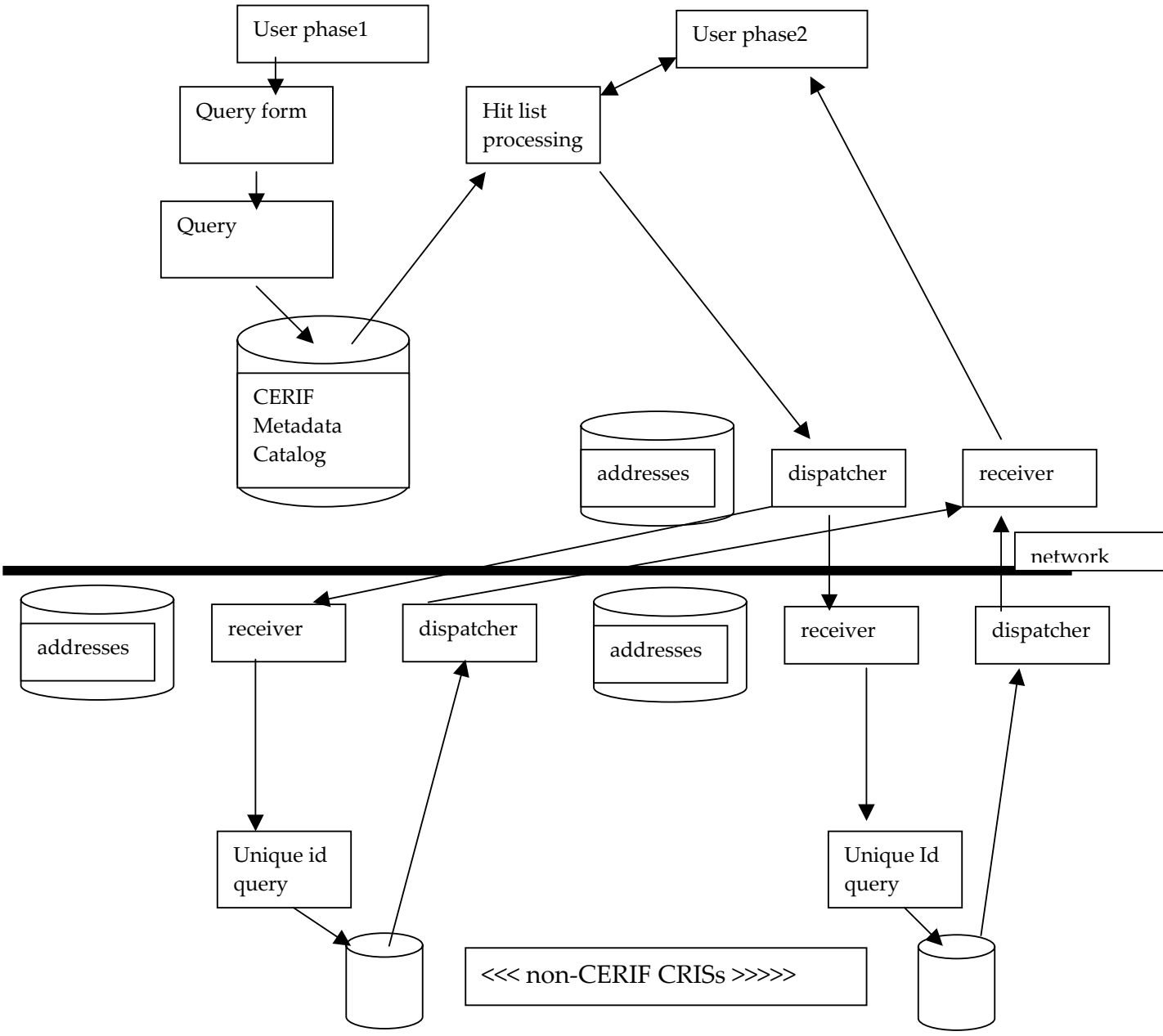
(a) simple query on union catalog (which may be centralised or replicated)

(b) possibly not all required entities and attributes in catalog

(c) effort to populate catalog; requires converter at each host to supply CERIF metadata

4.2 Catalog plus pull (ERGO 2++)

4.2.1 Architecture



4.2.2 Description

In addition to the Catalog-only model, this architecture allows a subsequent access to all hosts with hits in the CERIF metadata catalog to collect the detailed information the hosts are willing to supply. There is no further selection by attribute value nor projection of attributes, everything is 'pulled' if it relates to a hit record in the catalog. Ideally, the hosts convert to CERIF export model to provide uniformity but this is not mandatory.

4.2.3 Metadata

This architecture uses associative descriptive metadata (CERIF metadata catalog) and navigational metadata for host addresses.

4.2.4 Process

The hit list is edited by the end-user and then sent to the dispatcher which sends to each host the Unique Ids (primary key values) of the selected records for which further information is demanded. The host sends the answers back via dispatcher to receiver and thence to the user. No attempt is made to make homogeneous this detailed heterogeneous information which may have different character sets, language, syntax and semantics. Conversion to CERIF export model could be advantageous for integration by the end-user.

4.2.5 Advantages and Disadvantages

(a) advantage of simplicity as for catalog-only architecture

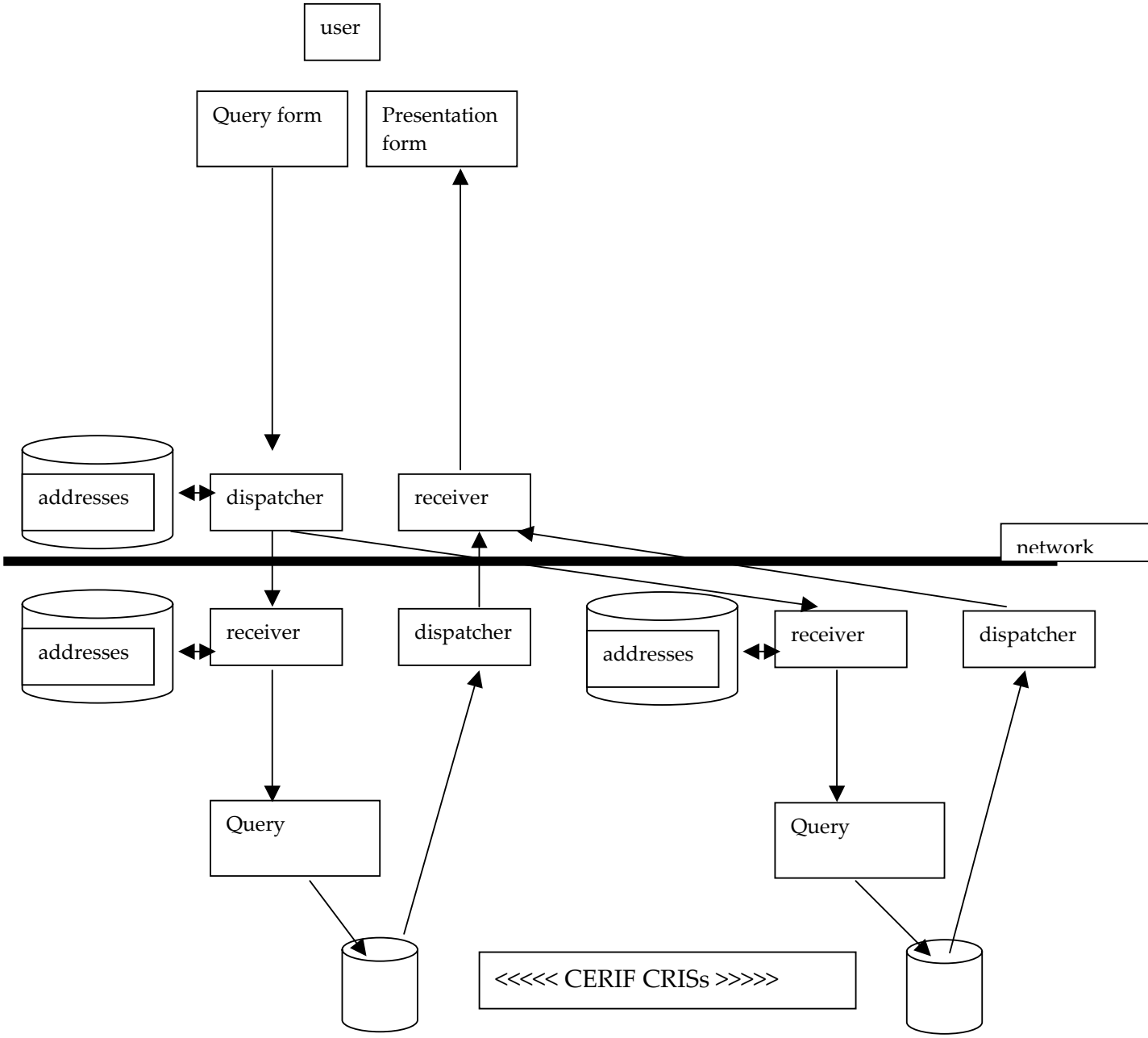
(b) advantage of additional information provision

(c) disadvantage that additional information is heterogeneous (unless converted to CERIF export data model)

(d) disadvantage of hosts having to maintain entries representing their database content in the CERIF metadata catalog

5. FULL CERIF

5.1 Architecture



5.2 Description

This architecture relies on the existence at each host of a full CERIF model database, either as the host database itself or a version of the host database converted to full CERIF model. This provides a completely homogeneous solution which is very simple.

5.3 Metadata

This architecture uses navigational metadata for host addresses. No other metadata is required as homogeneity is achieved through the full CERIF model.

5.4 Process

The process is straightforward; through a webform the end-user queries (knowing the CERIF schema) and using normal distributed database technology the query is passed to all hosts; the answers are all in CERIF form so integration is automatic.

5.5 Advantages and Disadvantages

(a) very simple and easy to use for the end-user

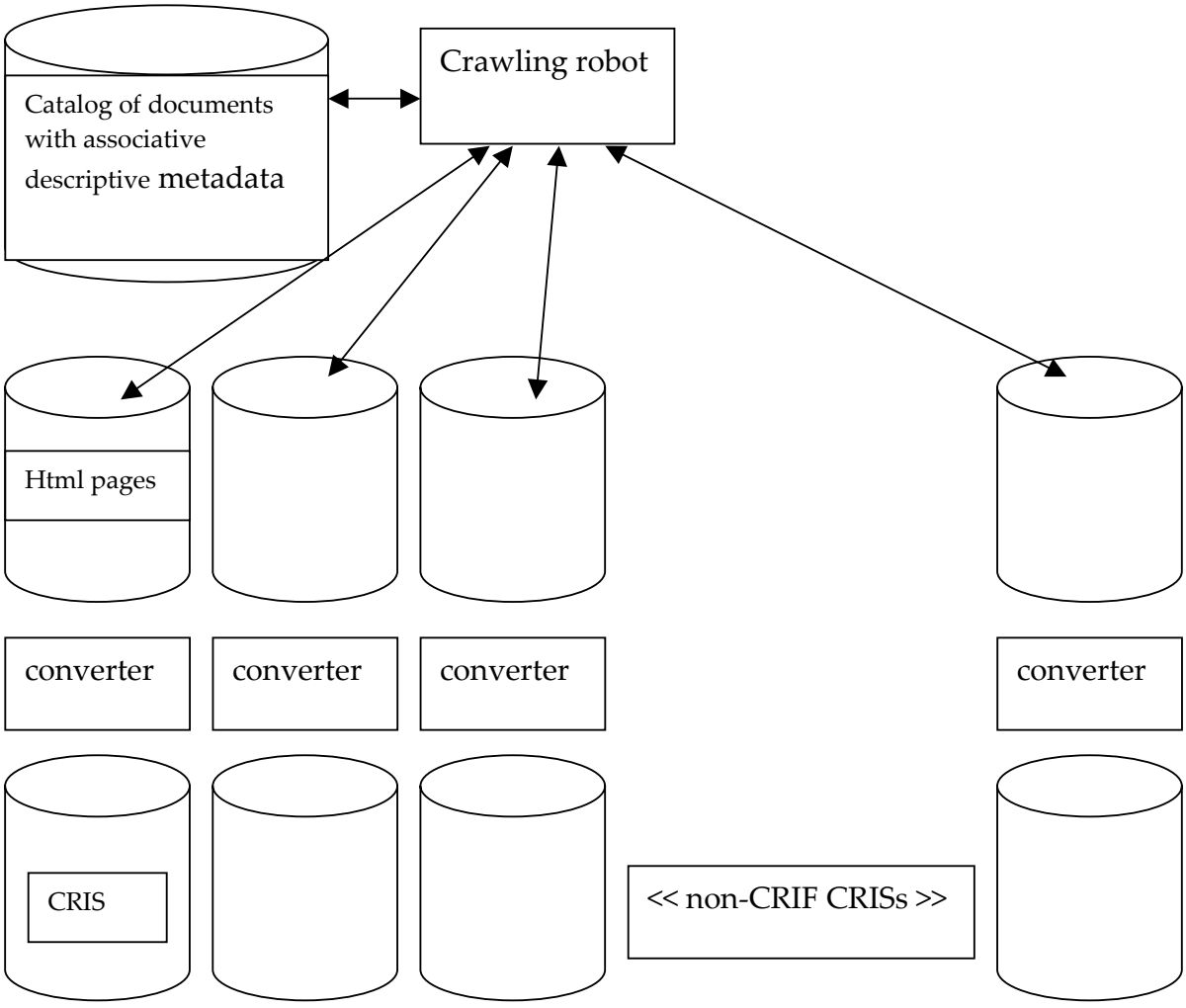
(b) each host has to either run a full CERIF model database or provide a full CERIF model version of the host database

6. HARVESTING

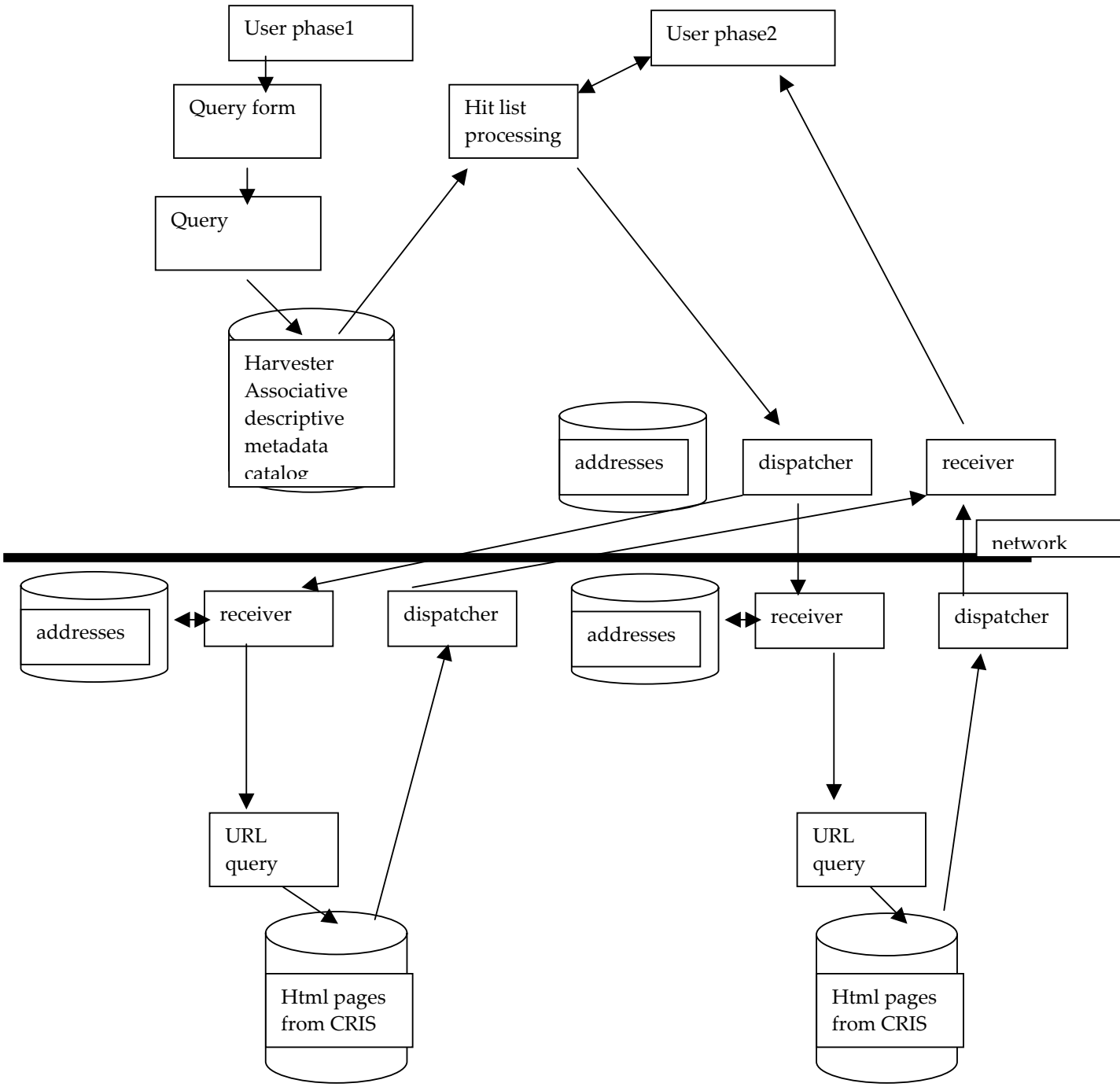
The concept of harvesting information from the whole WWW has been introduced. The power of modern search robots (to construct the catalog) and search engines (to search the catalog) is quite remarkable. However, much information is unavailable to harvesting being hidden in databases which may have a webform for query but which do not expose their information on webpages. Furthermore, the search robots usually take around 2 weeks to search the web and so the catalog is not up-to-date. A CRIS harvesting system should be more specific than, say, Google; this implies it searches only URLs known to be entrances to CRISs. The architecture (with catalog and reference to more detailed information) is not unlike the Catalog architecture of ERGO, but ERGO is based on structured data searchable under entity and attribute.

6.1 Architecture

6.1.1 Construction Phase



6.1.2 Search Phase



6.2 Description

This architecture relies on fast indexing of visible web pages by a search robot generating an associative descriptive metadata catalog which is then searched by the user; hits are followed up with a click on the URL to make available the detailed original web page.

6.3 Metadata

This architecture uses associative descriptive metadata in a catalog derived from the search robot and used by the search engine and navigational metadata for host addresses.

6.4 Process

First a search robot traverses the web; it may be instructed to search for certain terms but more likely is general. It constructs an associative descriptive metadata catalog as it goes, usually one entry per web page visited; the catalog also includes navigational metadata: the URL of the webpage indexed. This implies that any CRIS has to provide a set of web pages replicating the data in the CRIS to make it available to the search robot. Techniques are emerging to make structured or semi-structured databases visible to robots but there is no generally accepted technique yet.

The user then queries the catalog, and for every hit (meeting the search term(s)) receives a metadata record; clicking on the navigational metadata (URL in the metadata record) provides the original webpage.

6.5 Advantages and Disadvantages

(a) the host has to provide a copy of the database as webpages to be available to the search robot and subsequent accesses based on clicks from URL of metadata

(b) the query is based on existence of term(s); constraining by entity or attribute is not possible (without sophisticated xml form processing)

(c) the results are unstructured and one page at a time (click on URL in metadata catalog to see page); this inhibits statistical processing or report generation

(d) easy to implement and maintain (although may be ~2 weeks out of date)

(e) familiar interface for many WWW users

7. CONCLUSION

Clearly a full CERIF architecture provides maximum homogeneity and ease of use. However, it requires all hosts either to have their CRIS in CERIF or to provide a CERIF compatible version of their CRIS and make that version available to the federation system.

CERIF can, with advantage, be used as the canonical model for conversion from other CRISs when integrating using either remote or local wrapper techniques. It reduces the (n*(n-1)) interconversion problem to (n), where n is the number of participating CRISs.

CERIF metadata provides structured query capability in the catalog model(s), distinguishing this technique from harvesting.

Under any efficient architecture, CERIF remains the core technology for homogeneous access to heterogeneous CRISs.

ANNEX1 Metadata

Jeffery, K G: 'Metadata': in Brinkkemper,J; Lindencrona,E; Solvberg,A: 'Information Systems Engineering' Springer Verlag, London 2000. ISBN 1-85233-317-0.

ANNEX 2: CERIF

http://www.eurocris.org/ 'about' ==> 'taskgroups' ==> 'cerif'

ANNEX3: EXIRPTS Protocol

Naldi F, Jeffery K G, Bordogna G, Lay J O, Vannini-Parenti I
A Distributed Architecture to Provide Uniform Access to Pre-Existing Independent, Heterogeneous Information Systems
RAL Report 92-003

ANNEX 4: ERGO Final Report (Architectures and Cost-Benefits)

Remarks to the ERGO working group :

1.) Please check this draft, it is not yet short enough.
Mark things for deletion/shortening.

2.) Please check also, if this draft is clear enough,
and do suggest modifications where appropriate.

Kind regards, Bernd Niessen

DRAFT

European Research Gateways Organisation (ERGO)

Final Report to the INNOVATION PROGRAMME COMMITTEE

(Primary Part)

Version 1.0

17/2/05 13:56

ERGO (15) Final.doc

Table of Contents

8. 1. EXECUTIVE SUMMARY

A clear need for assisted uniform access to heterogeneous research databases is confirmed as predicted in the Terms of Reference.

On the basis of in-depth analysis of

- the perceived user requirements
- existing partial solutions both on organisational and technical levels
- existing information technology
- existing considerations on legal and political aspects

the working group considered

- the spectrum of possible models
- within the spectrum, variants of the optimal (catalogue) model.

The proposed solution combines elegant use of modern information technology (without special research components required) responding to the user needs and information providers' interest, with particular attention being paid to political, security, privacy aspects, as well as de iure or de facto standards. Other aspects, such as copyright and legal, have been considered in the light of EC directives.

The working group recommends an ERGO solution according to ERGO Proposal 2 , details see in chap. 6, which costs about 3.4 MECU and should be available to users in Member States about two years after the favourable project decision by the IPC, i.e. 13 months after project start, hence by 1999-03-30. This project will cover ERGO Phase 1, lasting for three years.

In addition to the above costs, 5 staff*days will have to be invested by the Database Providers for each database accessed via ERGO.

9. 2. INTRODUCTION

As directed by the Innovation Programme's work programme, to :

- facilitate access to national scientific and technical information services;
- identify the main existing information sources and to evaluate access possibilities and the potential for the utilisation of these sources at European level;
- develop the network bridges and instruments permitting access, on a harmonised basis, to these sources from other countries in the Community;

it is intended to set up a project on how to launch and operate gateways to Member States R&D databases. This is called ERGO, **E**uropean **R**esearch **G**ateways **O**rganisation.

10. 3. ERGO TARGET MARKET

10.1 3.0 Basic actors and their roles

The Target User (TU)

Role: The consumer of the information and, of course, the main target of the ERGO project.

The Data Provider (DP)

Role: Any Organisation (e.g. University, Public/Governmental Institution, Private Industry) with managerial/decisional autonomy, which funds research and collects and makes available information on research activities.

The Database Producer (DBP)

Role: The Organisation which makes available the information coming from the Data Provider.

The Data Nodes (DN)

Role: Each site which hosts the available full information. It is usually the same site where the original database. is hosted. It is possible that several Data Nodes are linked to the same Catalogue Node.

The Catalogue Nodes (CN)

Role: The sites which host one of the replicated copies of the ERGO catalogue, and makes the service available to the end user (Target User).

The Reference Node (RN)

Role: Technical point of reference for the ERGO system. Provides monitoring of the services, to ensure that updates sent from Catalogue Nodes (CN) are propagated to the other Catalogue Nodes.

The ERGO Management Board (EMB)

Role: Management and monitoring the strategic evolution of ERGO

The ERGO Task Force (TF)

Role: The technical management of the ERGO service

10.2 3.1 Target Users categories and user requirements.

Since 1995 the ERGO activities form part of the INNOVATION-Programme activities, and therefore the following users groups should be considered as first priority users [1]:

- Innovators in industry
- Researchers in research institutes
- Service and advisory organisations active in the field of research and technological development. (as Innovation Relay Centres)
- Information intermediaries
- Policy Makers and R&D Planners
- and to a certain extent, the Media

For these user groups, ERGO must meet the most relevant requirements within the limits of affordable effort required from database providers.

The ERGO **user interface** must be easy to use for both occasional and regular professional users. Therefore, standard Web browsers will be supported. It is desirable that the power needs of expert users are not overlooked.

During the initial Phase 1 of ERGO, the effort will be restricted to information on "research projects". The Council Decision refers to "national scientific and technical information services" as a whole.

10.3 3.2 Simple but viable charging

At present, many of the existing research databases in various countries are freely accessible. However, there are others in certain countries for which a fee is required from the users. ERGO will include a technical mechanism for charging the users (including monitoring). For full services, users will have to pay a flat subscription rate.

10.4 3.3 Reflections on the issues of data language and query language

ERGO will contain data in both the original database language and the commonly used reference language of the research community. Provided that funding will be available, ERGO will offer, as an extension, machine translation for query and research results in other languages of the European Union.

---

[1] The Council decision mentions as main target groups : enterprises, innovation supporting structures, technology providers and programme managers.

11. 4. SITUATION TODAY

11.1 4.1 Inventory

The working group analysed a number of existing inventories of relevant information sources. The following summary gives an outline on available candidates for an immediate integration into ERGO, which have been confirmed by the members of the working group.

Summary (### countries, ### DBs per country, types of DBs, etc.)

TO BE INSERTED

11.2 4.2 Relevant Projects

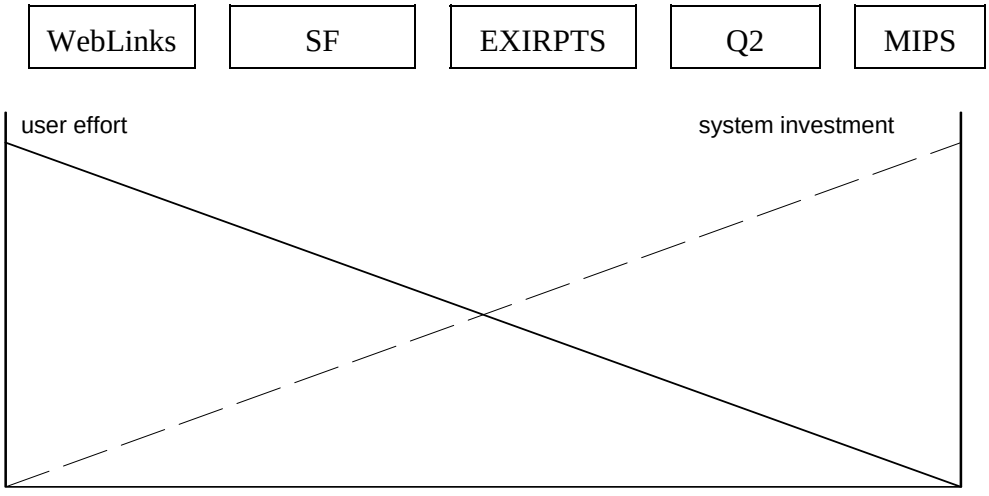
The working group identified many projects which had some component of relevance to ERGO*, including the complementary project on EuroCRIS Code of Best Practice*. These components were analysed and have been included where appropriate.

11.3 4.3 Relevant Existing Models: comparison

The working group considered both centralised single and distributed multiple models and recommended a distributed multiple approach. This allows for a high degree of European harmonisation without infringing the optimal solutions within individual Member States.

Five different architectural models have been analysed, reaching from a simple to implement Web Hyperlink approach leaving all the complexity to the user, over a catalogue based model solving the majority of the problems in an elegant way, till a full knowledge based model with high system investment and great user benefit.

The following diagram illustrates the models to building gateways evaluated by the working group.



The "**WebLinks**" model is based on a World Wide Web oriented solution characterised by hypertext links which help users to connect to other services. It leaves all the complexity of user

interfaces of these different services to the users. Complexity means different in user interfaces, database structures, coding schema, retrieval language, data language, display formats, etc. The different services are located in different countries. The only pre-condition for information access lies in the knowledge of the web location.

'SF' refers to the Search Fields Gateway solution, which has some advantages over the "WebLinks" Model but it does not provide true integration of information.

'Exirpts' refers to the catalogue based solution originally pioneered in the G7 Exirpts project. This prototype allows to assess current research projects funded by the participating institutions. Grace to its common catalogue, users benefit from a common structure, format and contents including one common user interface. Having been initialised by the top management level, acceptance by the participants was trivial. Technical implementation started immediately. The lesson we can learn from this project tells us that the most critical success factor is: Availability of data, which comes with regular updates from each participating host.

Only when all relevant updates become an essential part of the administrative procedures of each database producer, viability and success of such a service can be assured.

Q2 and MIPS are both Knowledge-Based Assisted Solutions (KBS).

Q2 model relies on the total and up-to-date emulation of terminal sessions to each and every database host being connected to this service. This concerns all levels of interfaces, such as telecommunication protocols, user interfaces, query languages, screen layout, structure of databases, coding, languages, etc.

The critical success factor lies in the complexity of the maintenance for continuous adaptation of all these interfaces to the database hosts. For these reasons, implementation of this model is limited to connecting only a small number of database hosts.

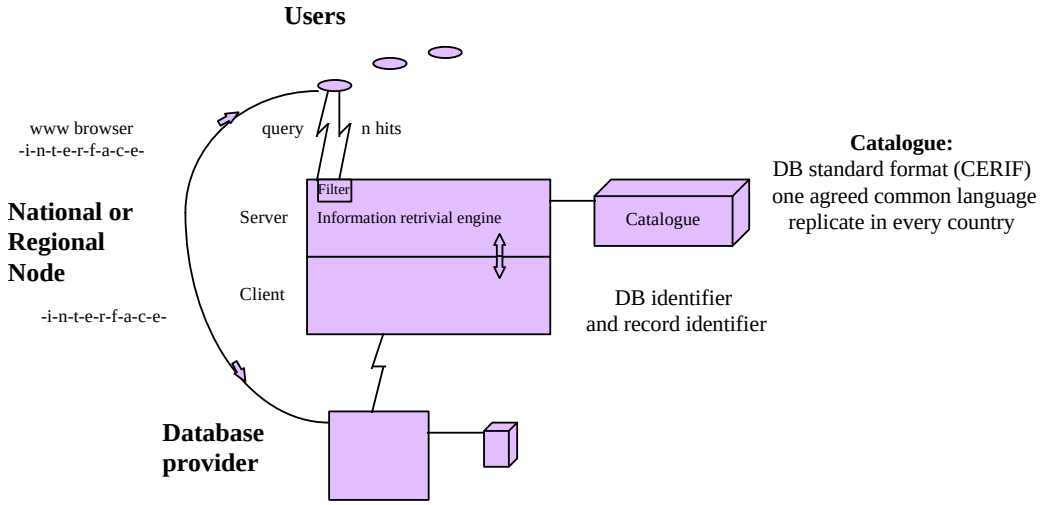
Q2 refers to the EC JRC ISPRA directed project. It should be noted that Q2 has problems when integrating more than a few target host databases.

'**MIPS**' refers to the 'Knowledge-Base Assisted Solution" demonstrated as a pilot system in the application domain of tourism by the ESPRIT MIPS project (ESPRIT project number 6542).

11.4 4.4 Recommendation of Catalogue Model

On the basis of cost / benefit, and taking into account non-functional requirements, the working group recommended the adoption of the catalogue model as being the basis for detailed proposed solutions. The advantages of this model are flexibility, performance, viability, adaptability. It combines the advantages of negotiated logical centralisation with physical distribution management and control.

11.4.1.1 The Catalogue-Based Model



Filter: The filter allows each Database Provider to decide what information is provided.

A system with the architecture represented in the diagram will handle requests according to the process outlined below:

a) The user interface is a WWW browser with forms capability (and possibly Java for enhanced end-user dialogue control and help)

b) The user enters the query on the form

c) The query is handled by the regional node

d) The query engine expects the query in a commonly agreed reference language and the query parameters presented in a standard form against the catalogue schema, such as CERIF

e) The number of matches in the catalogue is the number of hits

f) The query use the CERIF information in the catalogue as provided by the data provider; it is the decision of the data provider how much information to make available in the catalogue and how much to retain for (paying) access

g) The list of hits is presented to the end-user together with the rules and costs of requesting full information from the data providers; the user may abandon the query or agree to proceed (and pay)

h) For each hit there is a Unique Identifier (UID): this consists of the catalogue node id, the data provider id, the database id and the unique record number within that database.

i) The hits for a particular region are sent to that regional node as a message consisting of records of originating regional node id, query id, UID. It accepts UIDs, converts them into queries for the target data provider databases, wraps them in standard protocol for Email and sends them to the data provider databases

j) The data provider databases return (Email) to the regional node the documents (one per UID) as answers.

k) The regional node assembles the answers from its region into a folder of documents and send (Email) to originating node

l) The originating regional node accepts all the folders and assemble them into a complete portfolio for ending to the original querying user.

Clearly the regional node will maintain transaction logs. This will allow recovery actions to be taken on time-out, lack of delivery etc. The regional node can also implement access and security policies consistent with local legal requirement and the commercial requirements of their database providers. Regional nodes may implement value-added facilities including translation, additional query assistance, facilities for hyper linked multimedia display....

The regional nodes will require:

a) adequate computing power and data storage

b) excellent networking connections

c) skilled technical staff

d) agreements with regional data providers

Benefits

1) the architecture insulates the end-user from changes in data provider systems

2) the architecture provides for a free, open market system where data providers can choose what they offer at what price and users can choose for what they will pay at what price

3) the architecture insulates the data provider from many changes to the system; only provision to the catalogue of CERIF compatible records and Email access for query and answer to the database are required

4) the architecture allows regional nodes to negotiate with their data providers and allows local policies on payment and security

5) the architecture allows for development with more advanced techniques towards the "Knowledge-Base Assisted Model".

12. 5. ALTERNATIVES WITHIN THE CATALOGUE MODEL

The Working Group decided ERGO to be based upon a catalogue model. Within this proposed model, the Working Group elaborated a so-called "staged approach", in which several alternatives are proposed, where each alternative always is built on the inferior one. The different alternatives, explained in detail in the next session, therefore are called "Steps". 6 steps are provided, of which 3 are selected to be reasonable solutions. For these three alternatives, there is still the choice between different optional extensions.

There are two technical solutions, centralised and distributed. Each has separate advantages and disadvantages. There is clearly political advantage in a distributed solution, which, combined with the technical advantages, led the working group to the solution recommended.

One of the nodes will be managed by the Commission (e.g. CORDIS) and serve ERGO as the reference node.

12.1 5.0 Staged Approach ERGO Steps I to VI

The working group deliberately chose a series of evolutive steps providing progressively increased levels of information delivery and enduser comfort.

Each step builds upon the previous one but requires more resource for development.

12.2 5.1 Description of alternative ERGO solutions, following a staged approach

12.2.1 5.1.1. Basic principles of the proposed staged approach

The basic principles followed by the ERGO Working Group in designing a range of alternative concepts to facilitate access to R&D-databases in the Member States are:

- In Phase 1, ERGO will handle only R&D-projects databases.
- Start with a useful and affordable step in, which is already in line with the ERGO mandate.
- Each additional step brings additional value to end users.
- Keep the service / system flexible for enhancements, such as
  - multilingual versions
  - variable number of nodes
  - further types of databases