INTEREST

INTERoperation for Exploitation, Science and Technology

Keith G Jeffery
Director, IT &
International Strategy, STFC

keith.jeffery@stfc.ac.uk

Anne G S Asserson
Research Department
University of Bergen

anne.asserson@fa.uib.no

Authors

Keith G Jeffery
STFC-RAL

Anne Asserson
UiB

Structure

- Background
- The Hypothesis
  - Remote Wrapper
  - Local Wrapper
  - Catalog
  - Catalog Plus Pull (ERGO2++)
  - Full CERIF
  - Harvesting
- Conclusion

Background: GL

- Grey literature is important but is only a small component of the total research information environment and must be seen in context of the overall research process
- Grey literature is a product
- To understand the product need to have information on the sources and the process i.e. the research context

✂🡐 Do not try to obtain information through a 'fog' backwards from GL metadata

✂🡒 Get it moving forwards through the research process then much GL metadata derived directly and consistently

Background: Access

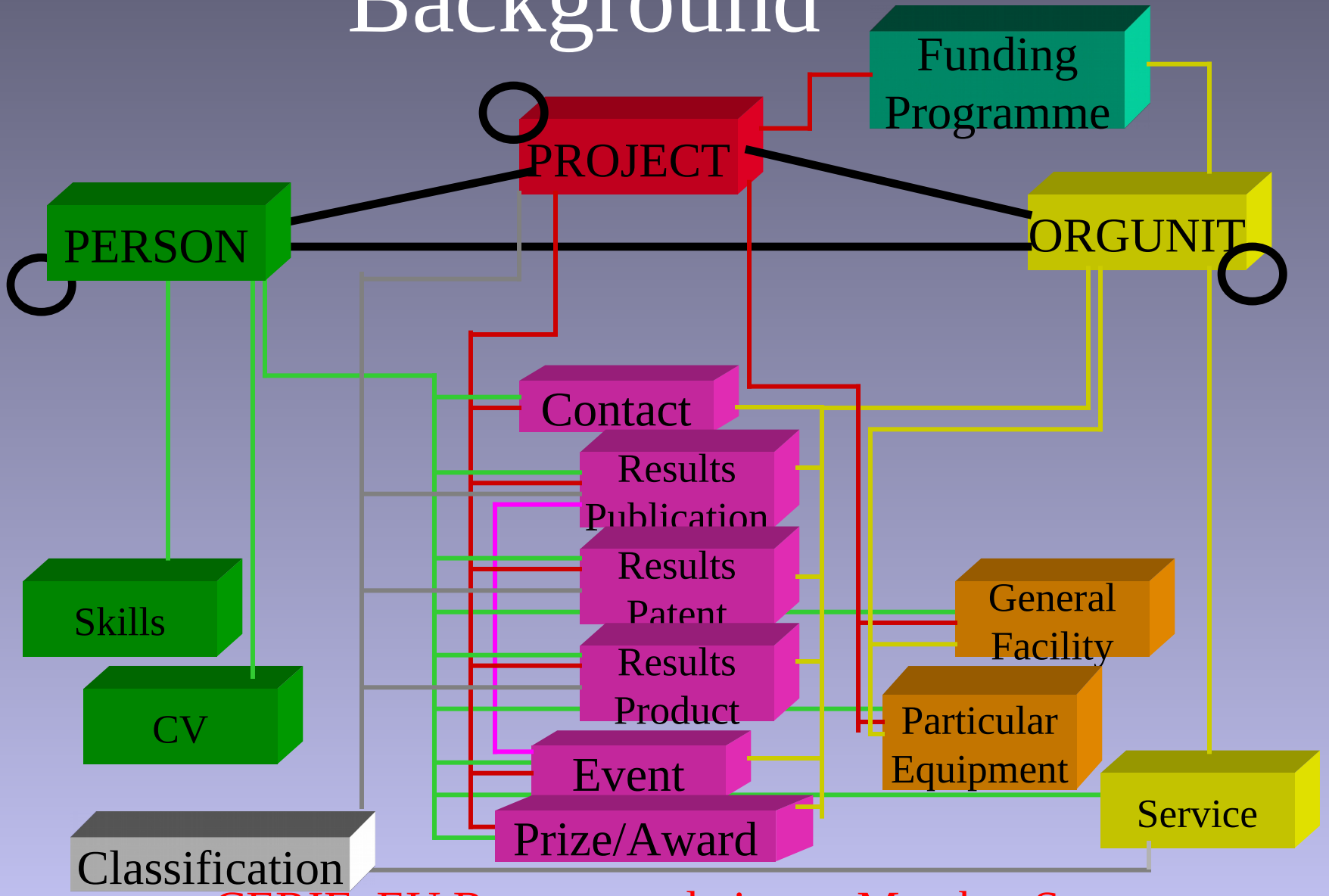
- Interoperation: homogeneous access to distributed heterogeneous information
  - Query against schema (of user)
  - Translation to other schemas (of sources)
  - Answer reconciled to original schema (of user)
  - If common interoperation format n interfaces
  - If not n(n-1) interfaces
- Utilise one common interoperation format
- [Character set, language, syntax, semantics]

- The alternative is ‘google-like’ where the end-user has to do the translations and reconciliations
- This does not scale

Background: Metadata

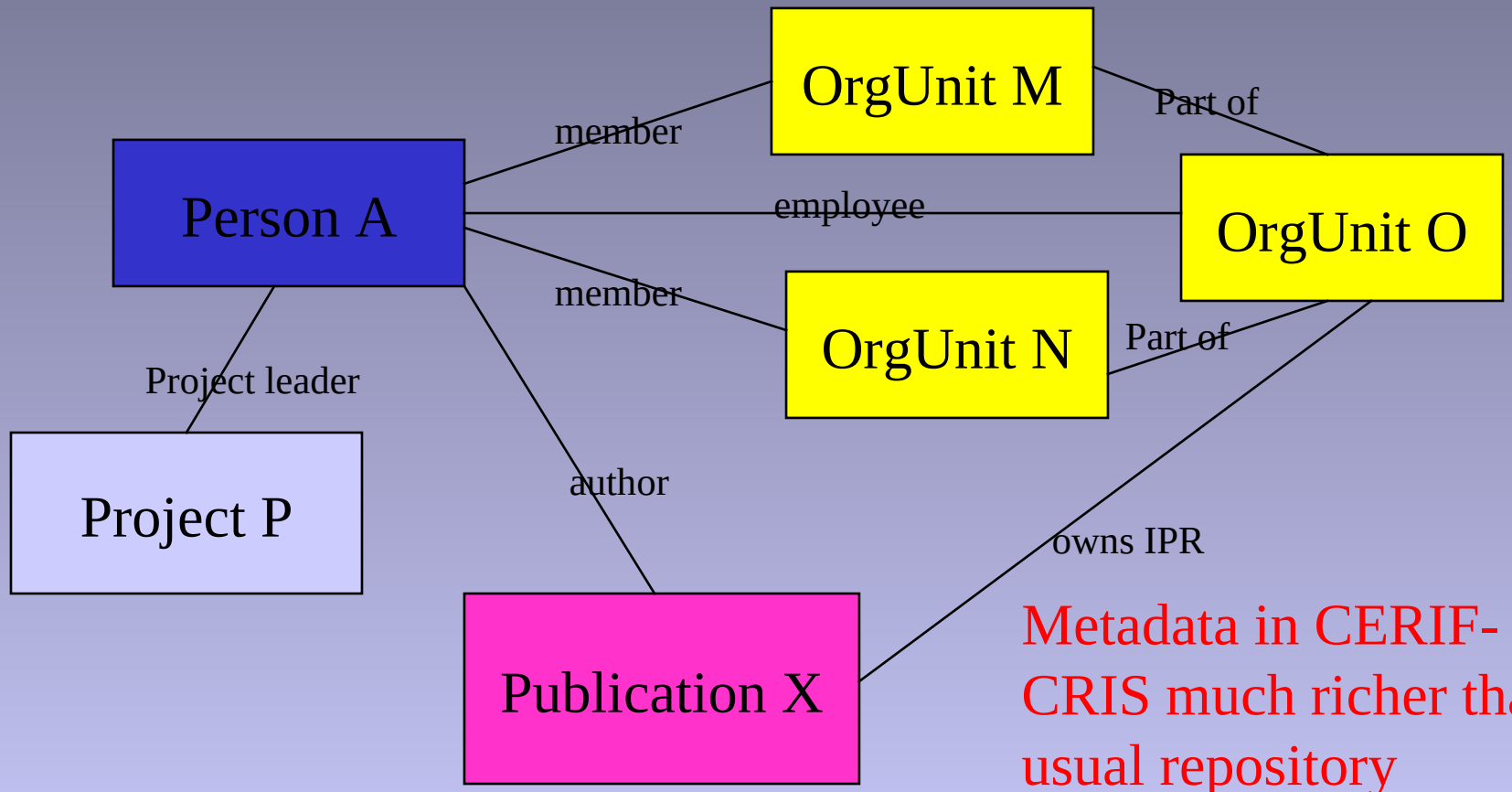
- Grey literature repositories can be interoperated without CERIF-CRIS using OAI-PMH and DC (OAISTER)
- Grey Literature Repositories provide better recall and relevance when interlinked via CERIF-CRIS – research context
- formal syntax, declared semantics
- Metadata
  - Schema, Navigational, Associative {descriptive, restrictive, supportive}
- The key to everything is quality metadata
  - input validation, query/retrieval, relationship linking, INTEROPERATION

Background



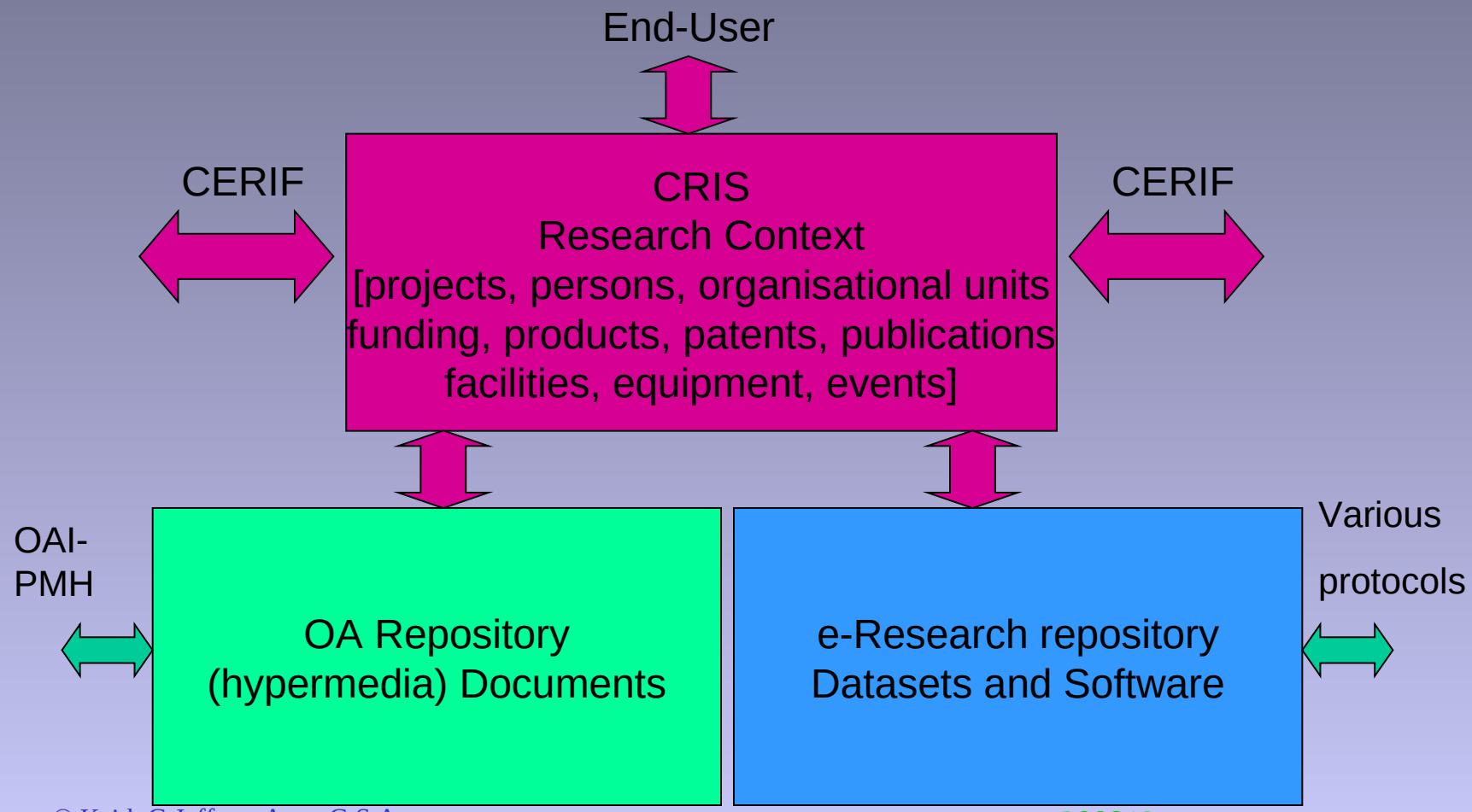
CERIF: EU Recommendation to Member States

Result Publication Instance Diagram

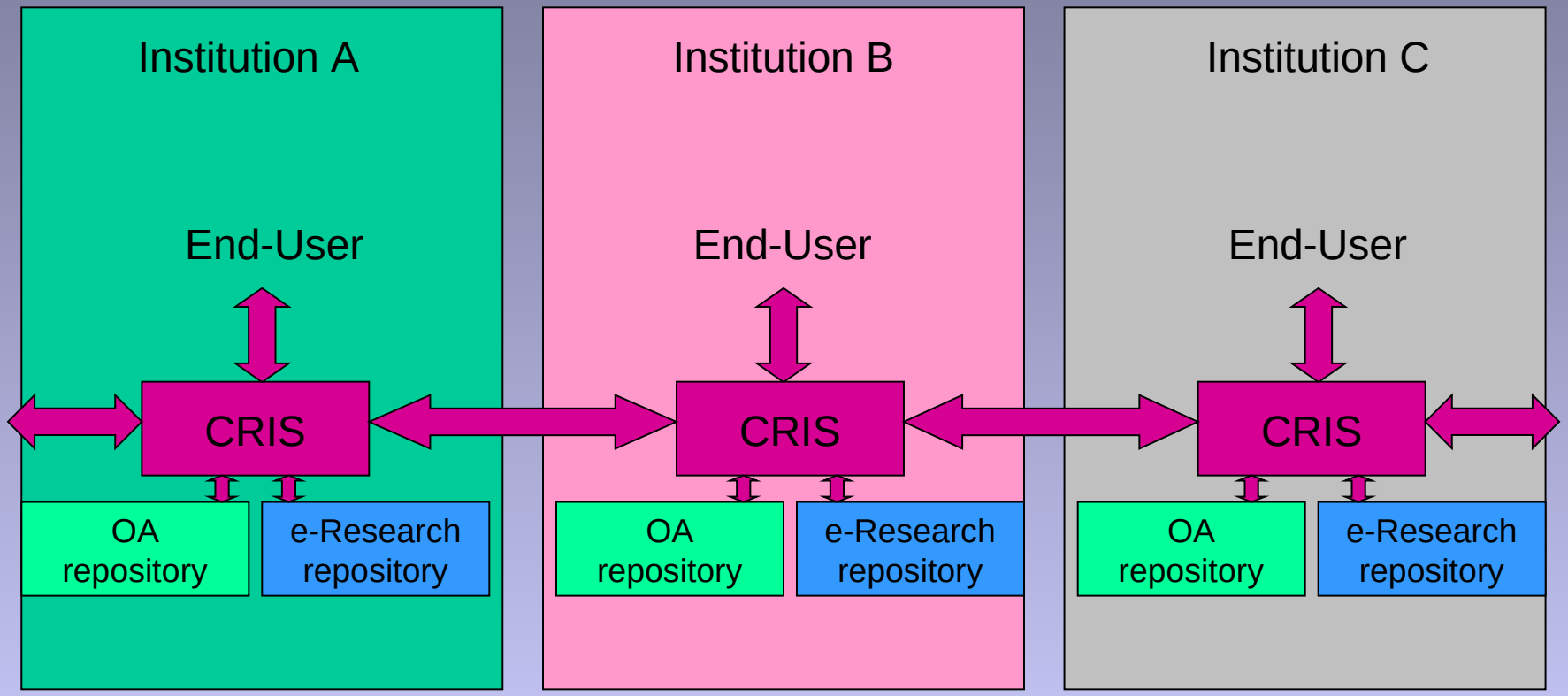


Metadata in CERIF-CRIS much richer than usual repository

CERIF- CRIS + Repositories at 1 institution



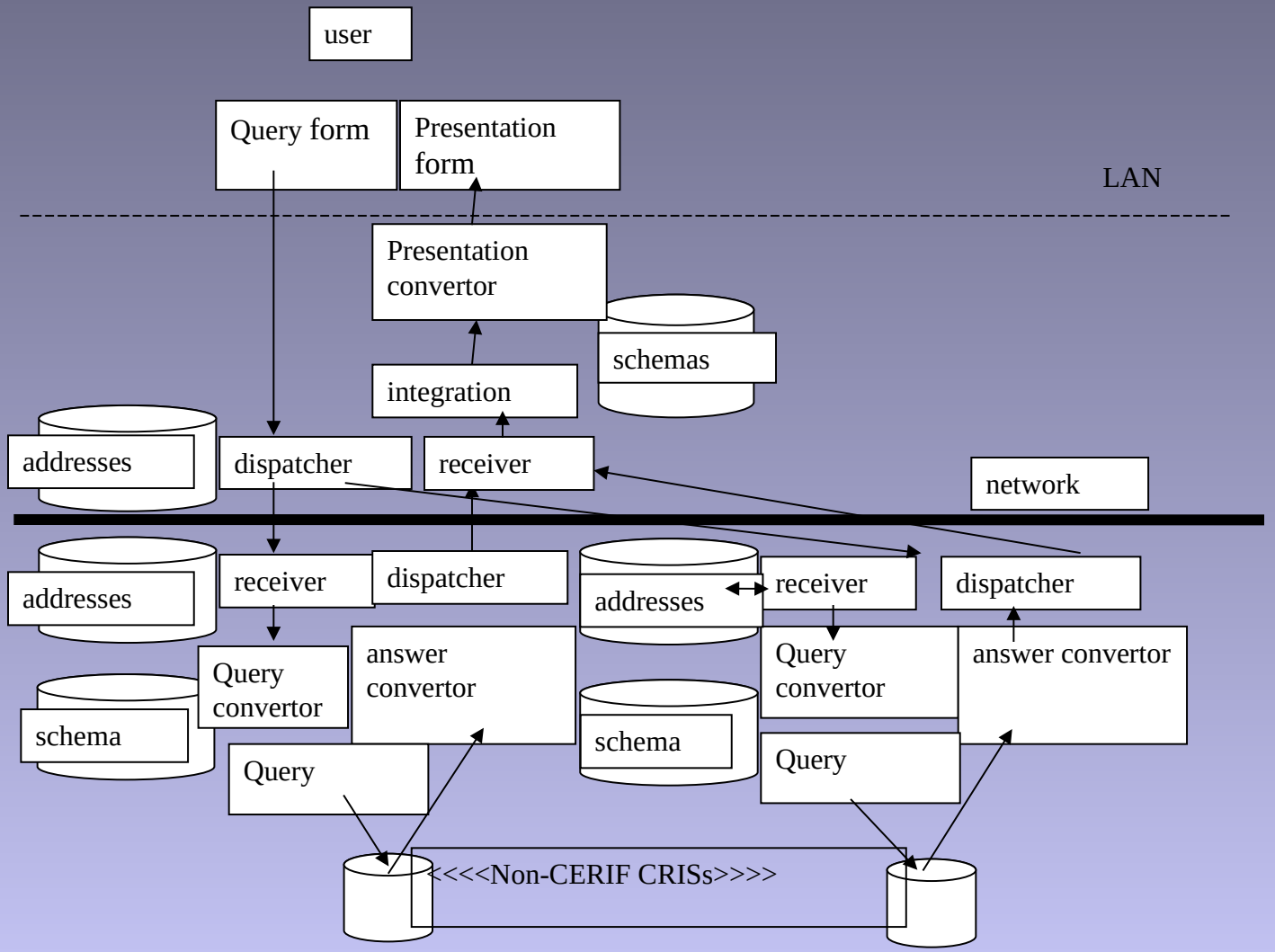
….and multiple institutions



Hypothesis

- Comparison of possible architectures for interoperation of grey repositories
  - (of publications or data and software)
- Leads inexorably to ===➔
- CERIF should be used either :
  - as the native storage format,
  - as the storage format of a derived data warehouse (transformed copy of the CRIS)
  - as the export format converted from the CRIS native format using a wrapper.

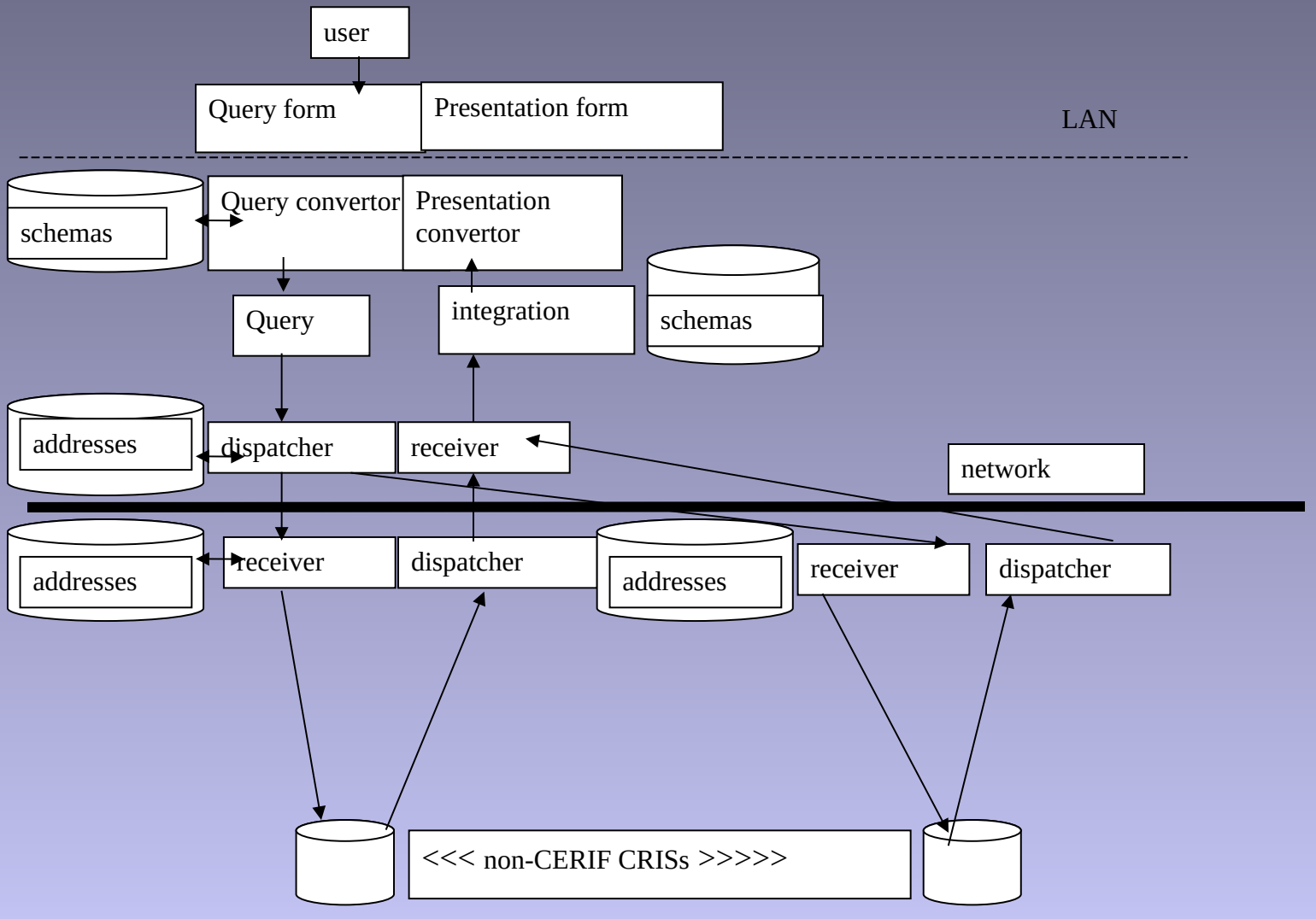
Remote Wrapper



Remote Wrapper

- the user needs only web browser and simple query form
- the host has to write query converter
- the host has to write answer (XML?) converter (to a specific XML DTD?)
- the query expressivity is very limited
- the user client has to write an integrator for the answers

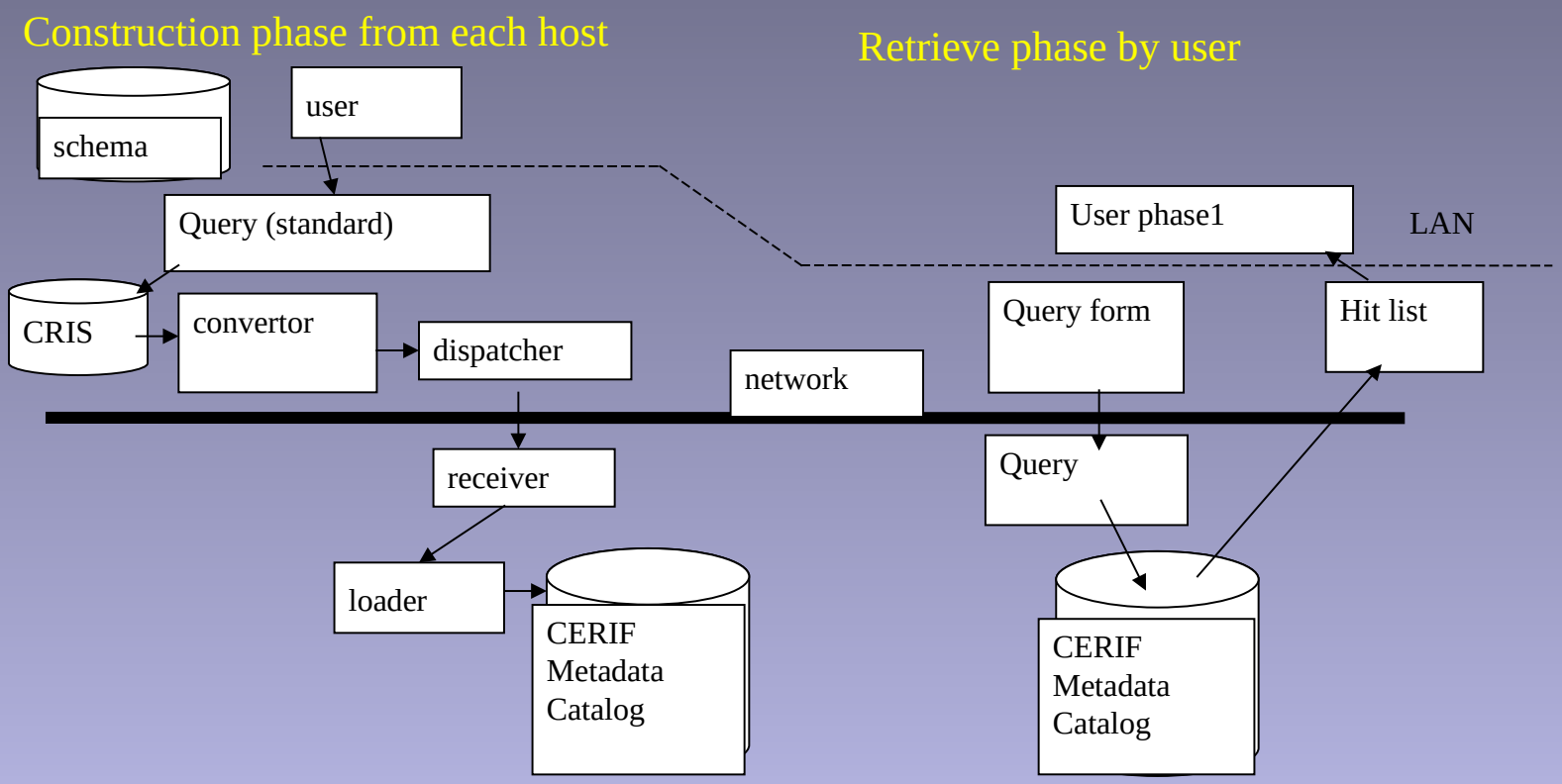
Local Wrapper



Local Wrapper

- each host has only to supply and update its schema to the client (all clients if there is not a central query server)
- each host has no software to provide except receiver and dispatcher
- the client (if it is a central service) has a very large workload
- if there is no central service then each client has to have all schemas supplied and updated
- the client software has to include a complex query refiner
- the client software has to include multiple complex query converters
- the client software has to include a complex answer integrator
- the client software has to include a presentation converter (complexity depends on specification of presentation required and complexity of the answer structure)

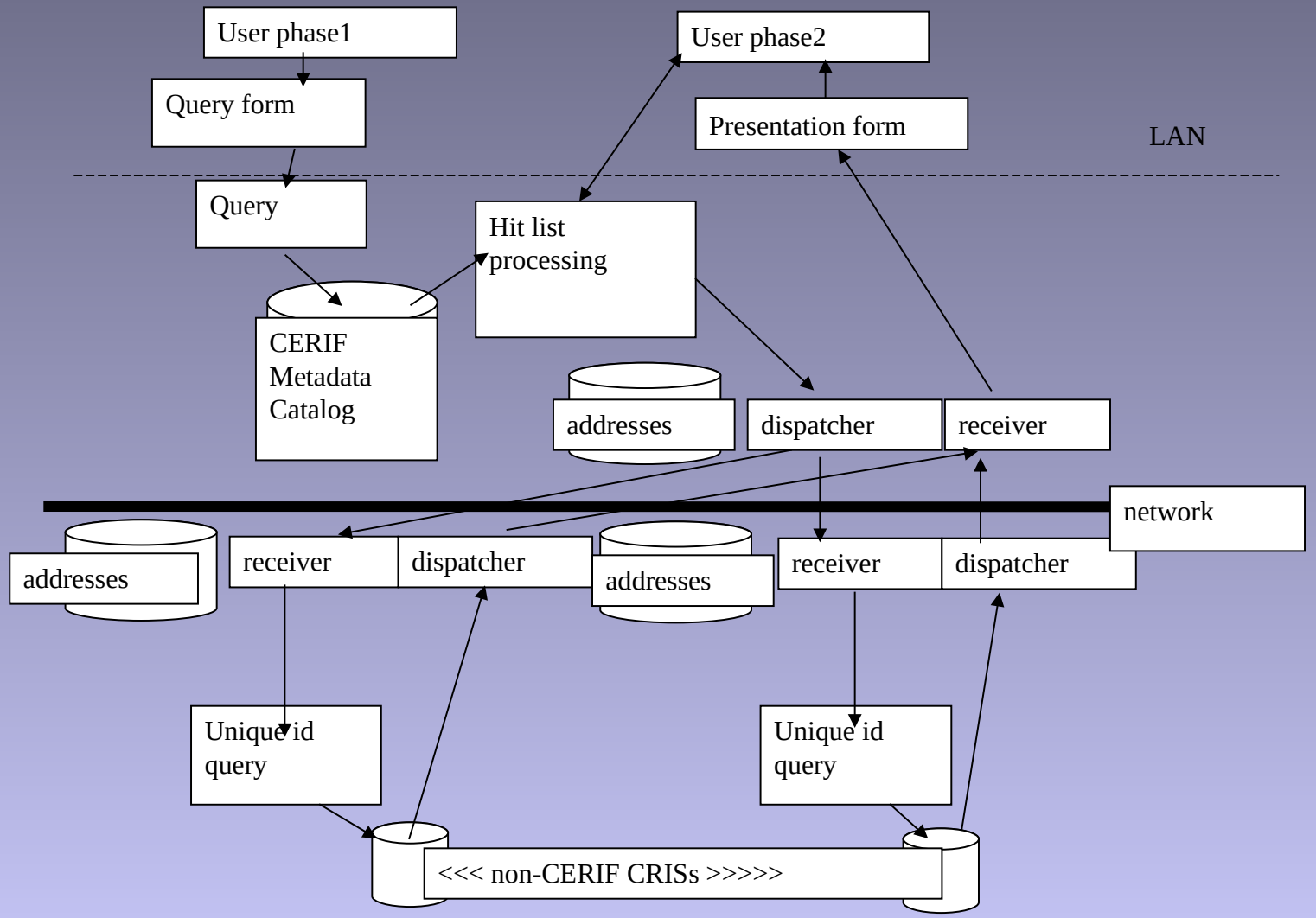
Catalog


Construction phase from each host
Retrieve phase by user
schema
user
Query (standard)
CRIS
convertor
dispatcher
network
receiver
loader
CERIF
Metadata
Catalog
User phase1
LAN
Query form
Hit list
Query
CERIF
Metadata
Catalog

Catalog

- simple query on union catalog (which may be centralised or replicated)
- possibly not all required entities and attributes in catalog
- effort to populate catalog; requires converter at each host to supply CERIF metadata

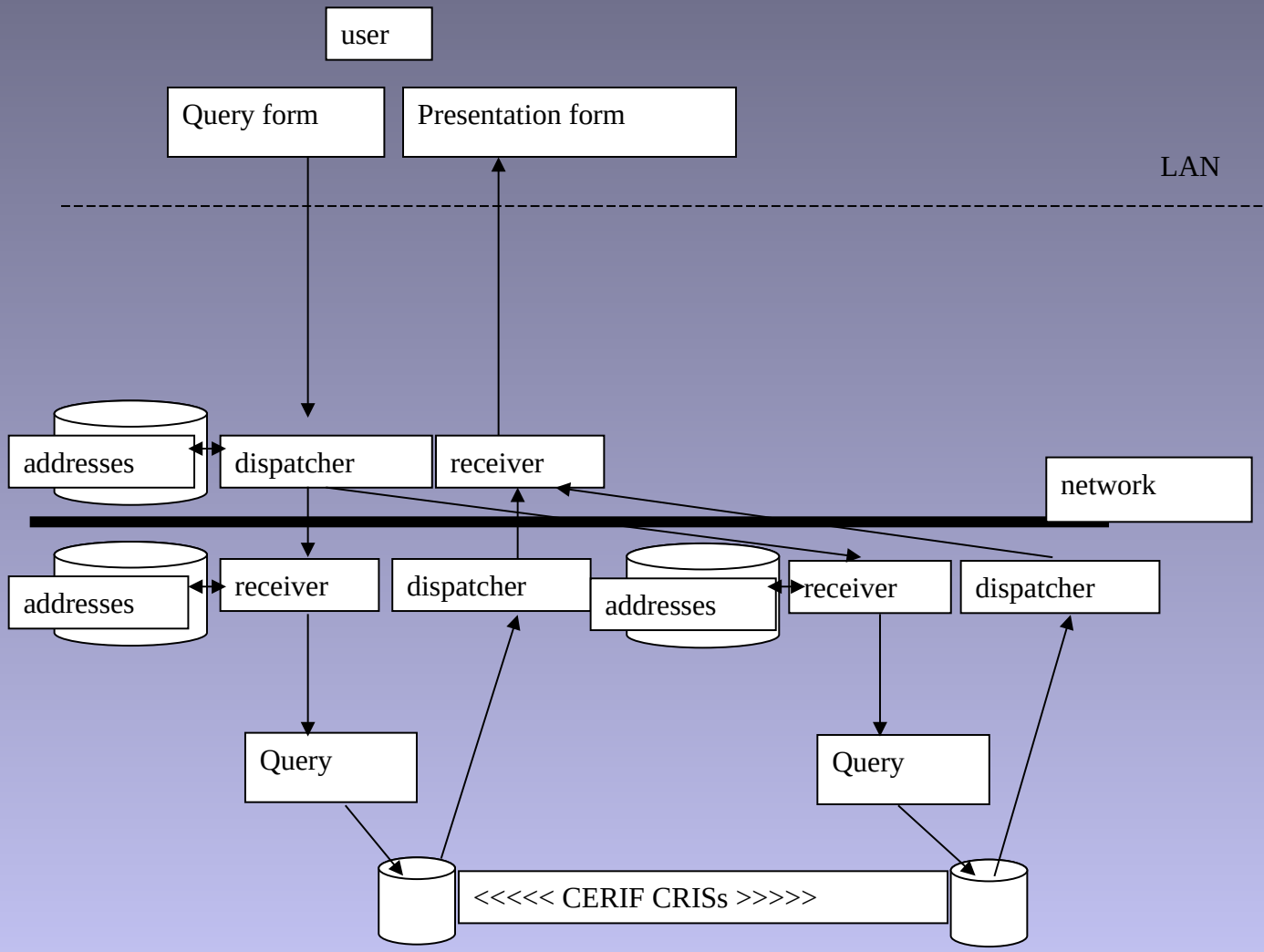
Catalog Plus Pull (ERGO2++)



Catalog Plus Pull (ERGO2++)

- advantage of simplicity as for catalog-only architecture
- advantage of additional information provision
- disadvantage that additional information is heterogeneous (unless converted to CERIF export data model)
- disadvantage of hosts having to maintain entries representing their database content in the CERIF metadata catalog

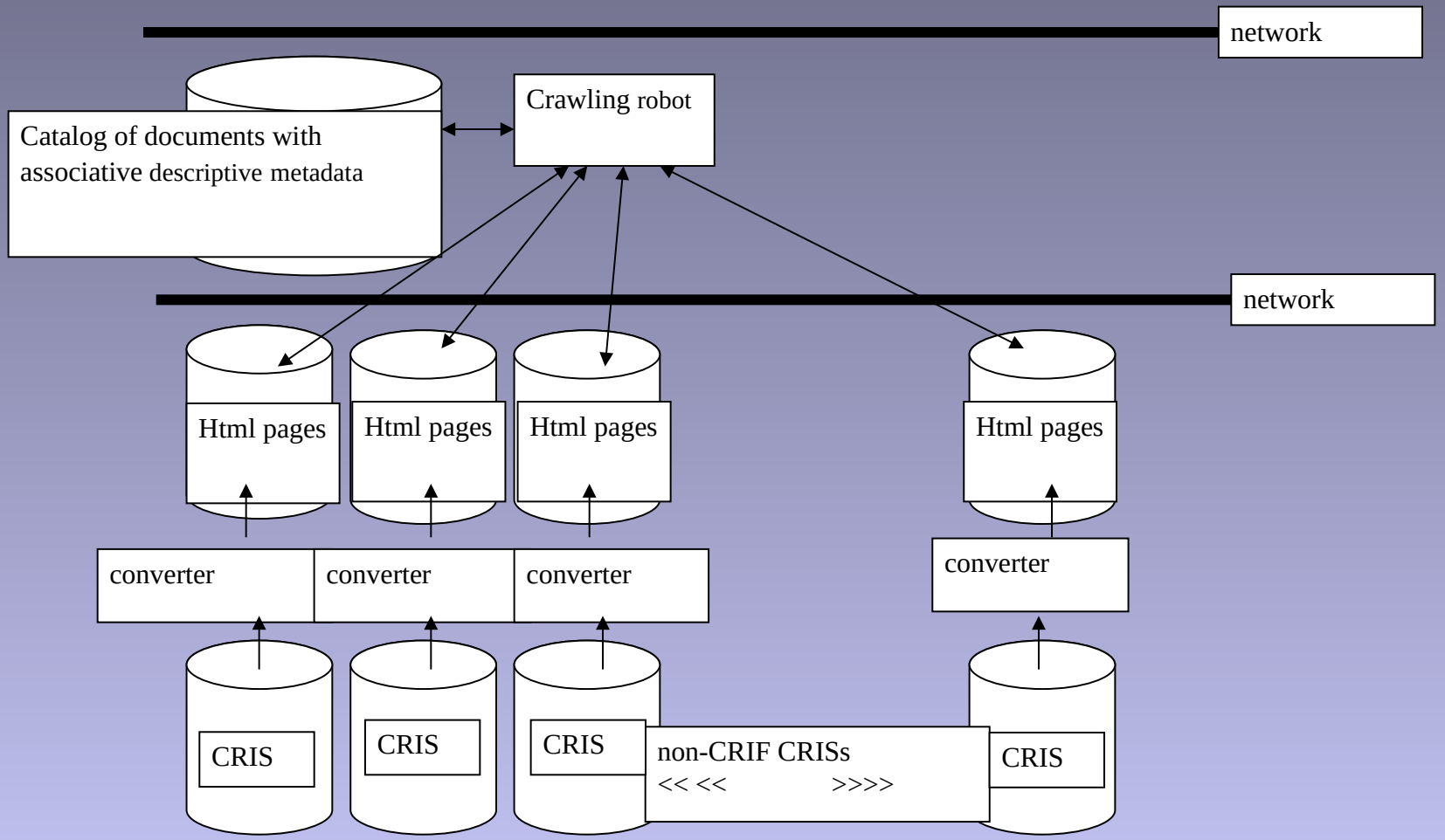
Full CERIF


user
Query form
Presentation form
LAN
addresses
dispatcher
receiver
network
addresses
receiver
dispatcher
addresses
receiver
dispatcher
Query
Query
<<<<< CERIF CRISs >>>>>

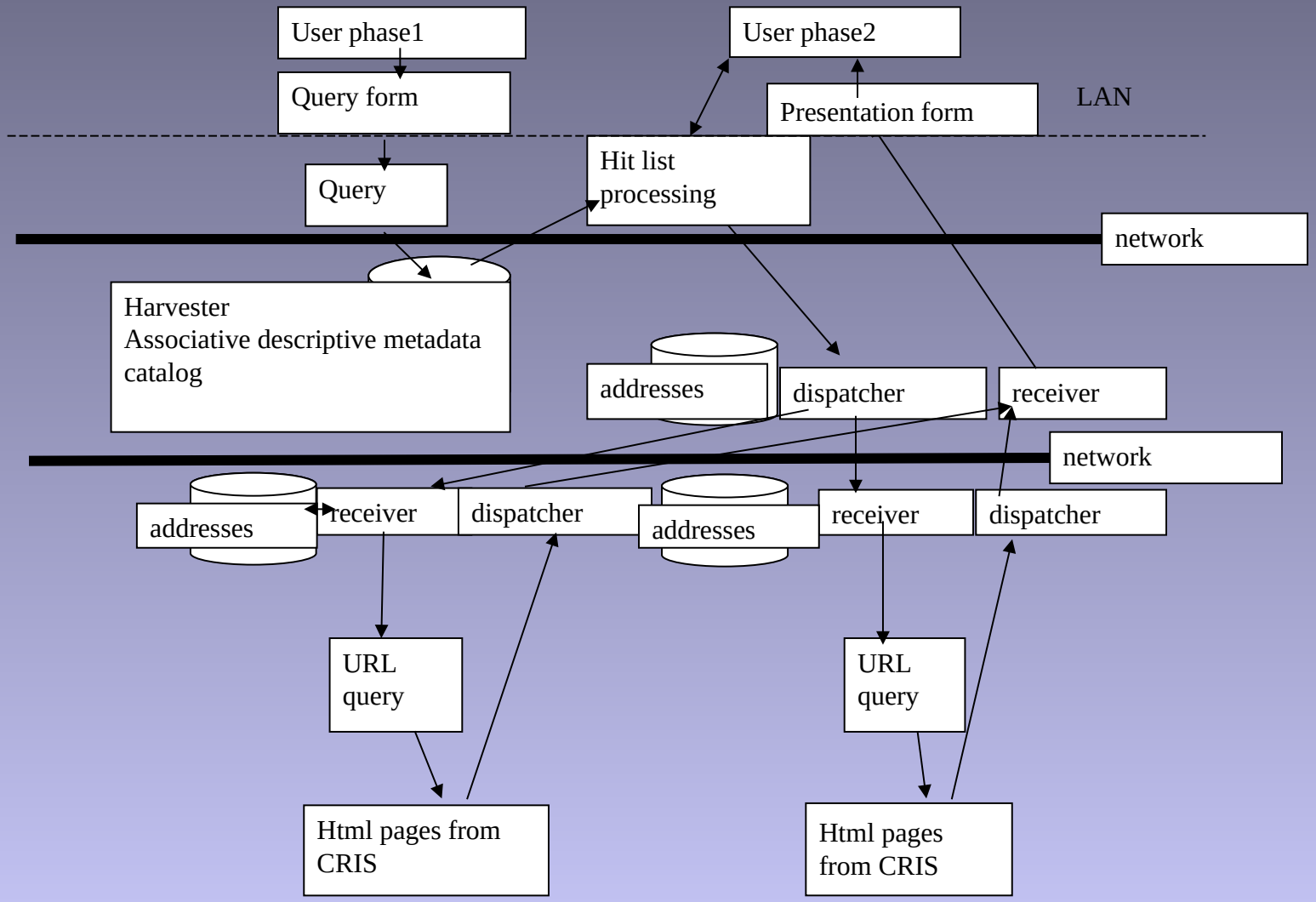
Full CERIF

- very simple and easy to use for the end-user
- each host has to either run a full CERIF model database or provide a full CERIF model version of the host database

Harvesting (construction phase)



Harvesting (search phase)



Harvesting

- 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.
- The query is based on existence of term(s); constraining by entity or attribute is not possible (without sophisticated xml form processing).
- 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.
- It is easy to implement and maintain (although the database may be ~2 weeks out of date) and has a familiar interface for many WWW users.

Conclusion

- ✓To interoperate grey repositories link to a CRIS
- ✓Best: Full CERIF architecture
- ✓Else: wrap CRIS to interoperate using CERIF