MOSAIC Shades of Grey

Realisation through Formalisation

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
- Proposed Architecture
  - Objects, Data and Metadata
  - Requirements
  - Architectural Solution
- Conclusion

Background: Mosaic / Grey

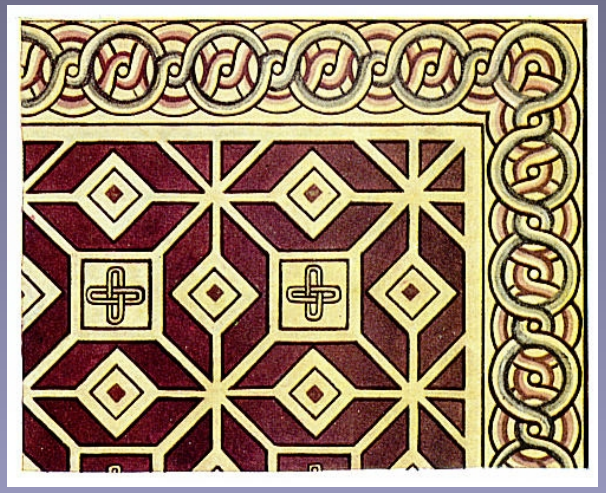
- Designed
  - For a purpose
- Formal structure
  - To improve access and understanding
- Composed
  - Of component pieces in structures
- Representation
  - Of something in the human mind
- Communicate
  - The idea to others
- Effort
  - To produce the grey objects and to provide the repository

Hypothesis

- A grey literature collection is much better collected, structured, catalogued, utilised and maintained within the context of a research environment (commonly known as e-Research or e-Science)
- which relies on CERIF-CRIS to provide
  - improved metadata for each GL object
  - contextual research information
  - access to other recorded research information
  - thus improving the integration and publicising of grey within the research scene.
- The key is
  - improved data collection,
  - improved interoperation
  - improved query relevance and recall
- all based on the formal syntax and declared semantics of a CERIF-CRIS.

Objects, Data and Metadata

- Conventional metadata in Grey repositories is insufficiently formal resulting in much end-user effort in
  - Input
  - browsing for retrieval
  - interoperation
- If metadata has formal syntax and declared semantics
  - Improved ease of data input
  - Ensuring quality of data
  - Providing automated retrieval with improved recall & relevance
  - Reliable automated interoperation



Requirements

- Input of grey object and its metadata
  - Pre-filling
  - Workflow (bite-sized chunks)
  - validation
- Retrieval of set of grey objects meeting criteria
  - Recall
  - Relevance
  - Homogeneous access over heterogeneous sources
- Subsequent processing
  - Count, sum, average ➔ graphics, modelling
- Relating to other information
  - To provide the end-user with the complete picture

Architectural Solution (1)

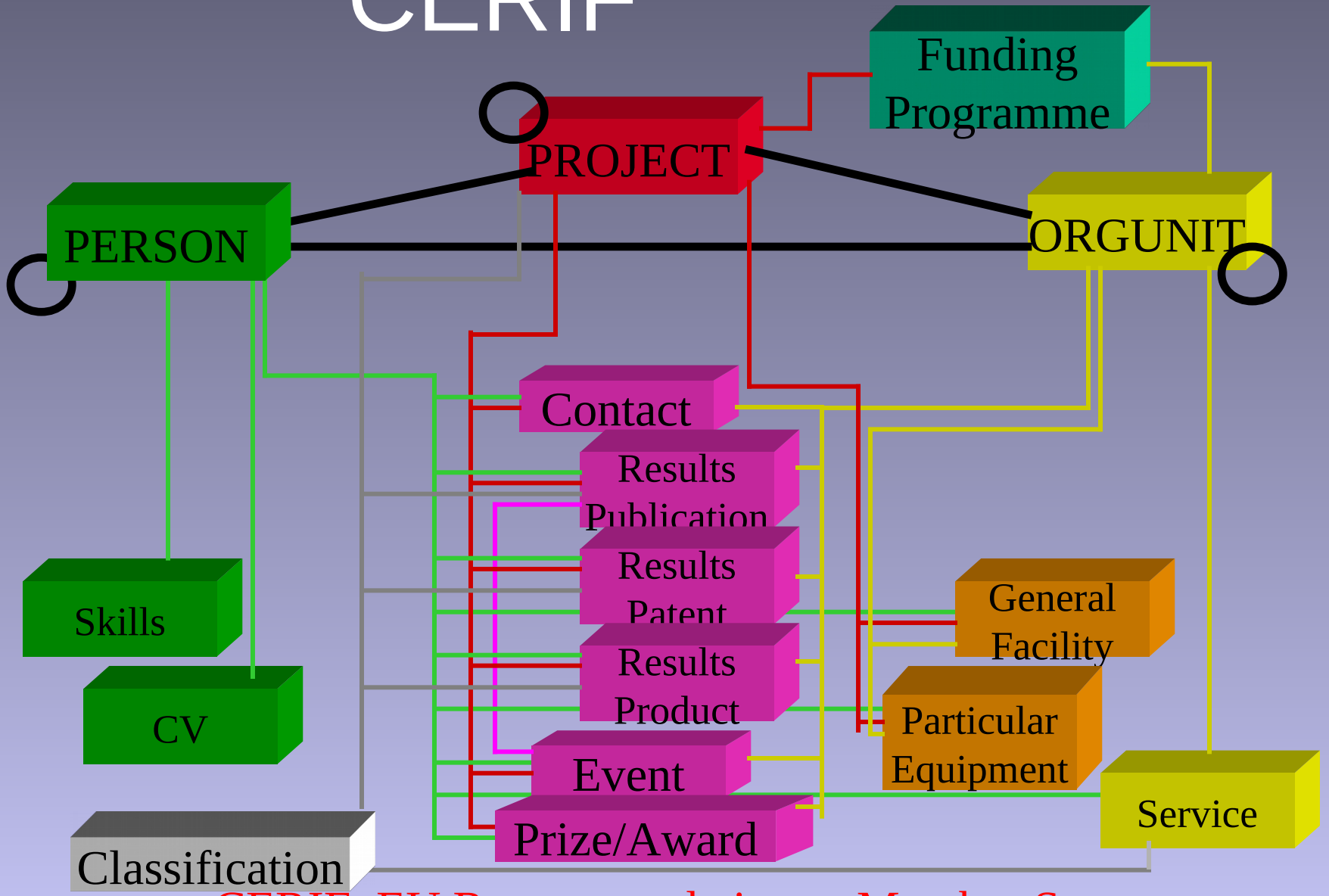
- same canonical schema;
- formal syntax and declared semantics;
- data for some purposes, metadata for others;
- linking relations between entities with date/.time stamp and role such that
  - the structure is articulated flexibly,
  - new entities can be added and related
  - links to external systems can be made

  using the same framework

Architectural Solution (2)

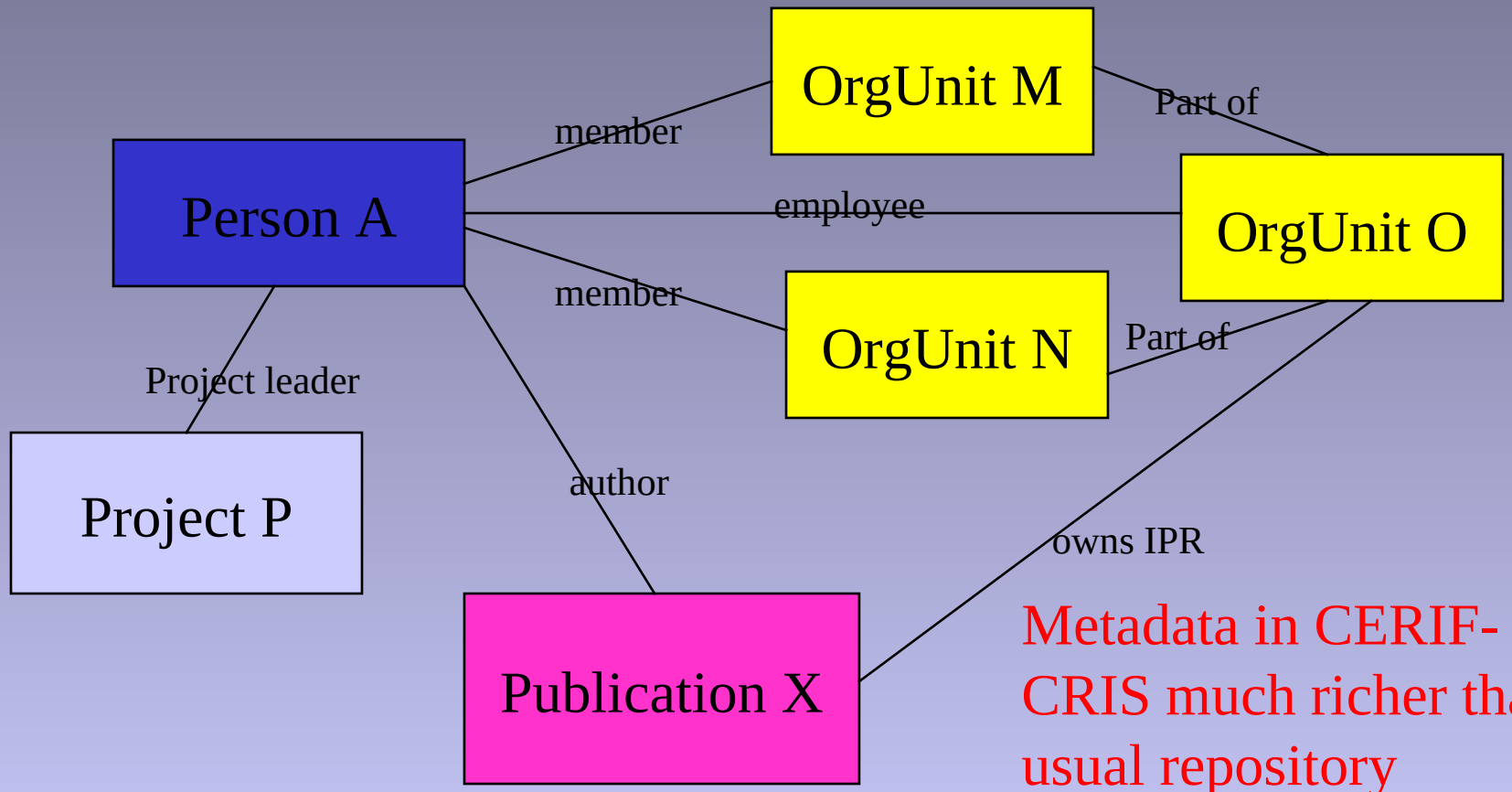
- the above provide an optimal base framework for the processing required including
  - input within a progressive workflow,
  - retrieval and reporting,
  - subsequent processing including statistical and graphical reports
  - interlinking to other systems both within and outside of the research organisation.

CERIF



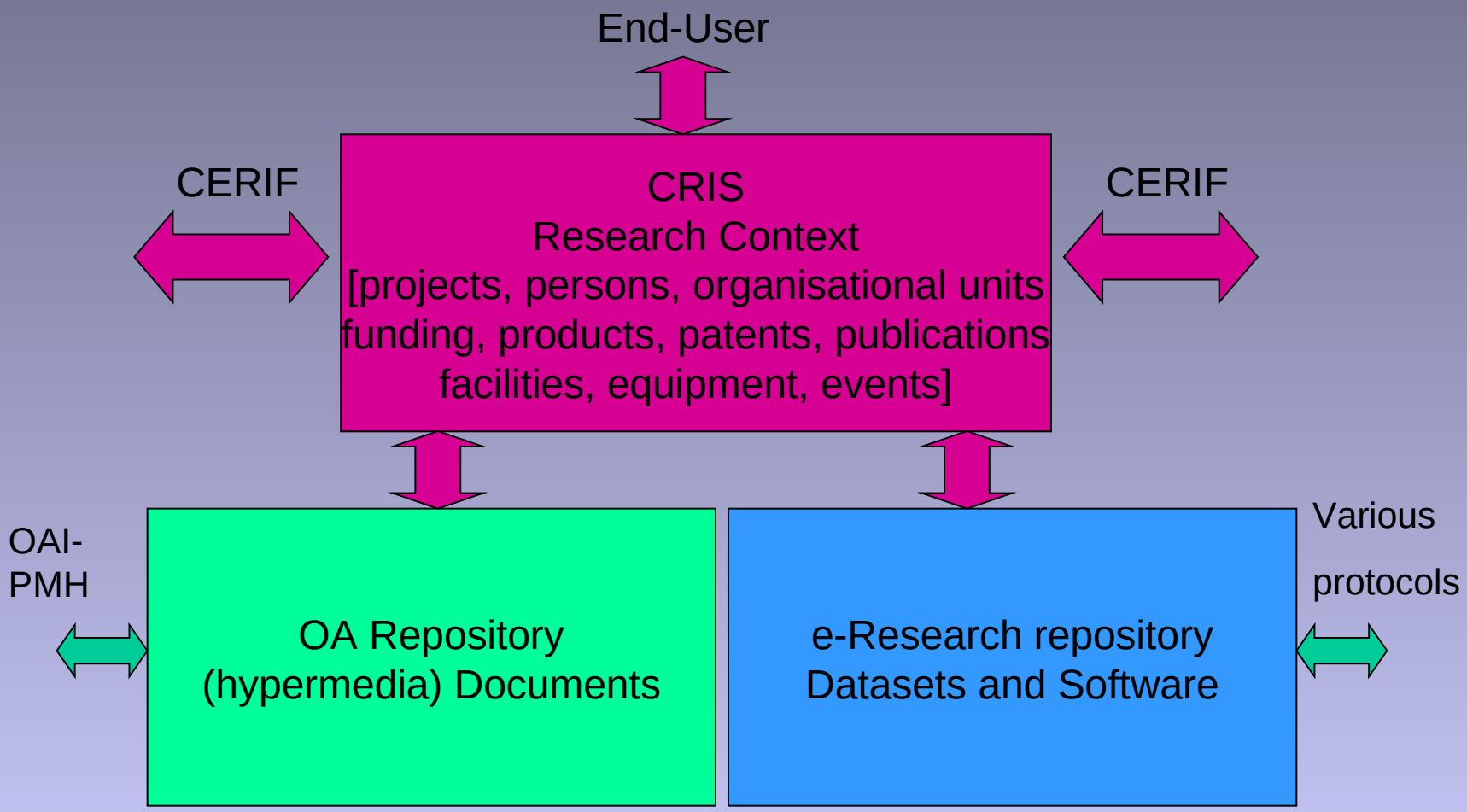
CERIF: EU Recommendation to Member States

Result Publication Instance Diagram

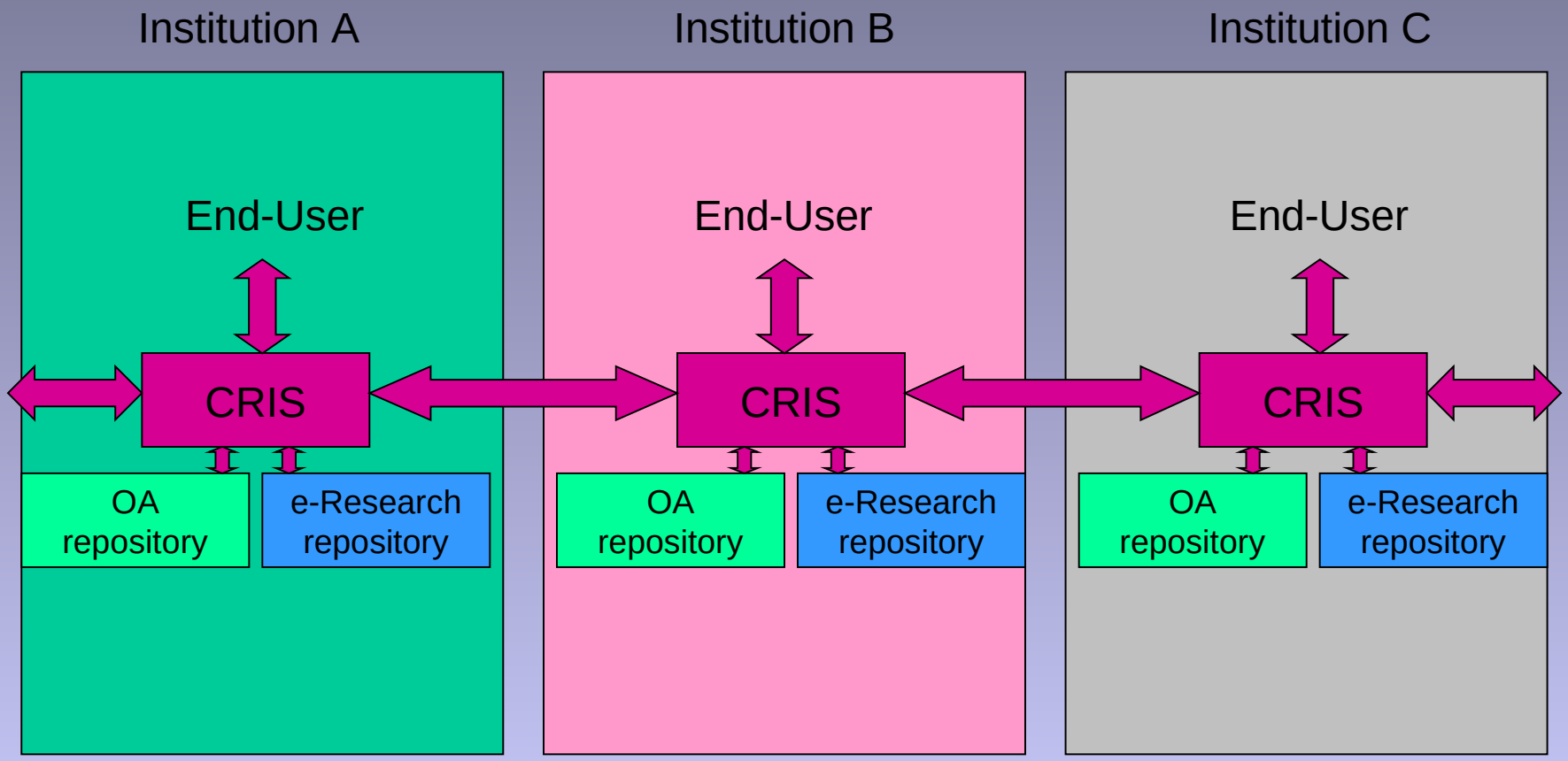


Metadata in CERIF-CRIS much richer than usual repository

CERIF- CRIS + Repositories at 1 institution



….and multiple institutions



Conclusion (1)

- The mosaic of grey literature is not yet revealed easily.
- Its complex patterns representing structures, and the beauty of the complete form are not recognised.
- This is because of
  - the heterogeneity of the sources,
  - the lack of a canonical schema either fo
    - storage/query/results management
    - interoperation over heterogeneous systems.

- Worse, existing sources use metadata schemas that
  - do not have sufficiently formal syntax
  - lack declared semantics
- both of which can be rectified by the use of CERIF.

Conclusion (2)

- The take-home message is clear: use CERIF as the canonical schema for grey literature.
- to accommodate legacy systems use a CERIF wrapper.
- This would mean that:
  - 1. query and retrieval provide better relevance and recall;
  - 2. data input quality is improved;
  - 3. systems can interoperate, to provide the end-user with a homogeneous view over heterogeneous distributed systems;
  - 4. statistical and graphical processing can be reliable;
  - 5. interoperation with other systems within and outwith the research organisation is facilitated.