# Auditing Grey in a CRIS Environment

Keith G Jeffery
Consultant

keith.jeffery@
keithgjefferyconsultants.co.uk

Keith G Jeffery Consultants

Anne Asserson
University Library
University of Bergen

anne.asserson@fa.uib.no

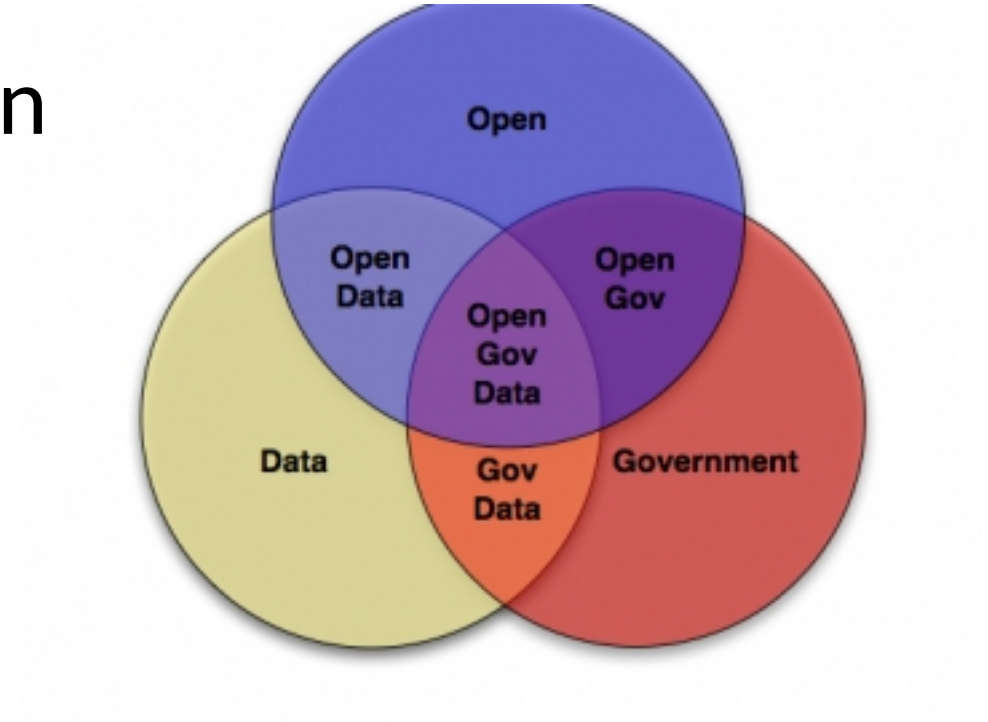
# Prologue

- Metadata and data
- Real world
- ‘library’ metadata: MARC, DC etc
- Key dependencies
  - Functional
  - Referential

- No AUDIT without QUALITY METADATA

# Structure

- Introduction
- Reliable Information
- Open Data
- ENGAGE
- Conclusion




engage
An Infrastructure for Open, Linked Governmental Data Provision towards Research Communities and Citizens

euroCRIS
Current Research Information Systems

# Introduction

- The vast majority of (research) information is grey
  - It is not peer reviewed scholarly publications
- We use information object to mean any digital grey object encoded in any format on any medium
  - Document, data file, video, software….
- Mechanisms are required to audit grey to assure quality
- We assert that audit of grey requires high quality metadata

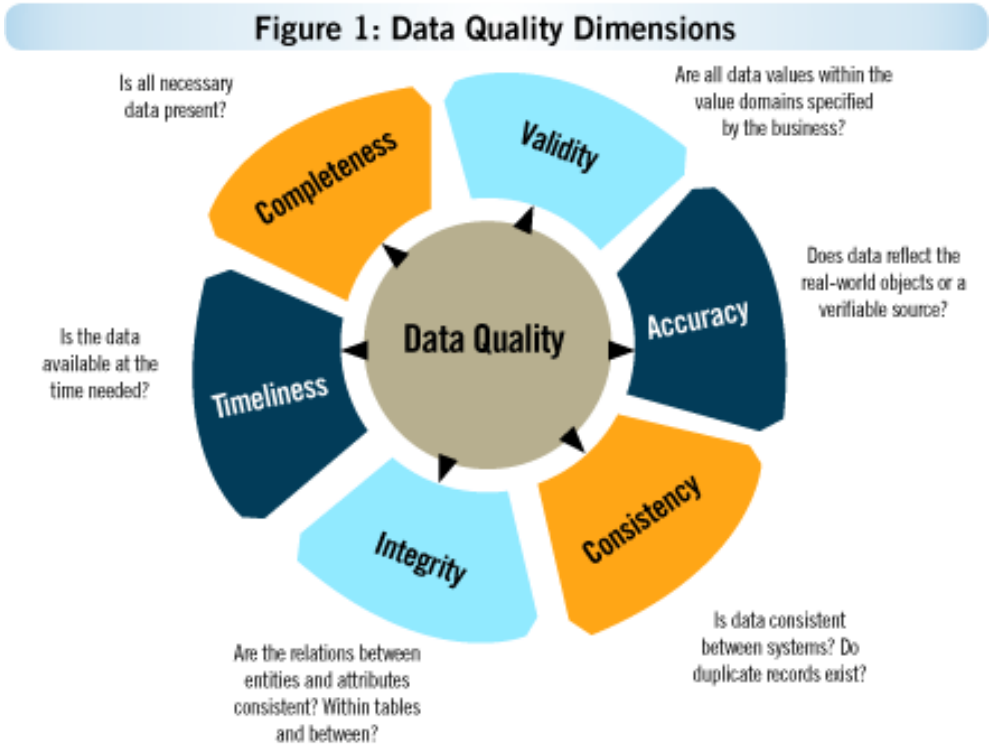
# Reliable Information

- Quality
  - Represents accurately world of interest
- Context
  - Environment within which collected – related entities
    - Persons, organisations, projects, funding, equipment, publications…..
- Availability
  - Persistence (preservation / curation)
  - Conditions of use (open access)

**We have to encode this as metadata for audit**

# Reliable Information: Quality

- Data integrity
  - Schema
  - Constraints
- Accuracy, precision
- Incomplete and inconsistent information
- Temporal validity
- Independent validation
  - Quality rating



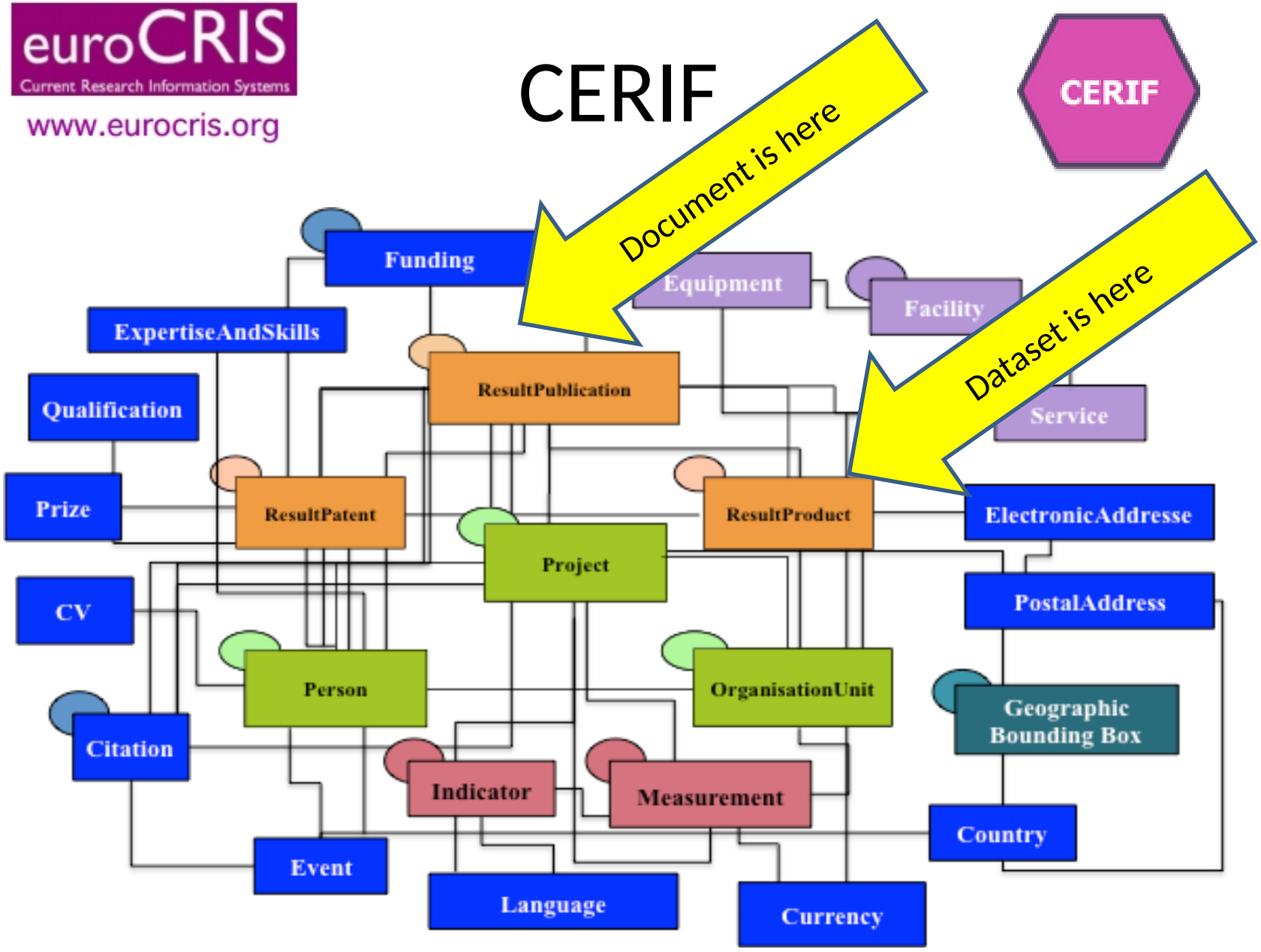

(With acknowledgements to FINETIK)

# Reliable Information: Context

- Related entities that give confidence that the information of interest is understood in context
- CERIF (Common European Research Information Format)
- EU Recommendation to member states
- Used in 42 countries
- National standard in 10
- Maintained, developed, promoted by euroCRIS (not for profit) www.eurocris.org

euroCRIS
Current Research Information Systems
www.eurocris.org

# CERIF



# Reliable Information: Availability

- Persistence
  - Media migration
    - Who can read a 7 inch floppy disk? Or a 3420 IBM tape?
  - Declared syntax and semantics
    - Machine readable AND machine understandable
  - Preservation of related software
    - Changing languages, compilers / interpreters
    - Changing operating environment (sequential, parallel, distributed, data dependencies)
    - Specifications
- Access
  - Open
  - Toll-free (conditions, licences)

# Open Data

- Semantic Web
- LOD: Linked Open Data
- RDF
  - Triples
  - Expressed as XML
- Metadata
  - DC
  - CKAN

- Most portals clickable lists of datasets
- Most datasets pdf or xls
  - Essentially documents
- Very little metadata
- Metadata ‘flat’ and poor
- Not linked to underlying research datasets

**Open data implies open access to any digital information object**

# Open Data

- Semantic Web
- LOD: Linked Open Data
- RDF
  - Triples
  - Expressed as XML
- Metadata
  - DC
  - CKAN

Poor 'flat' metadata

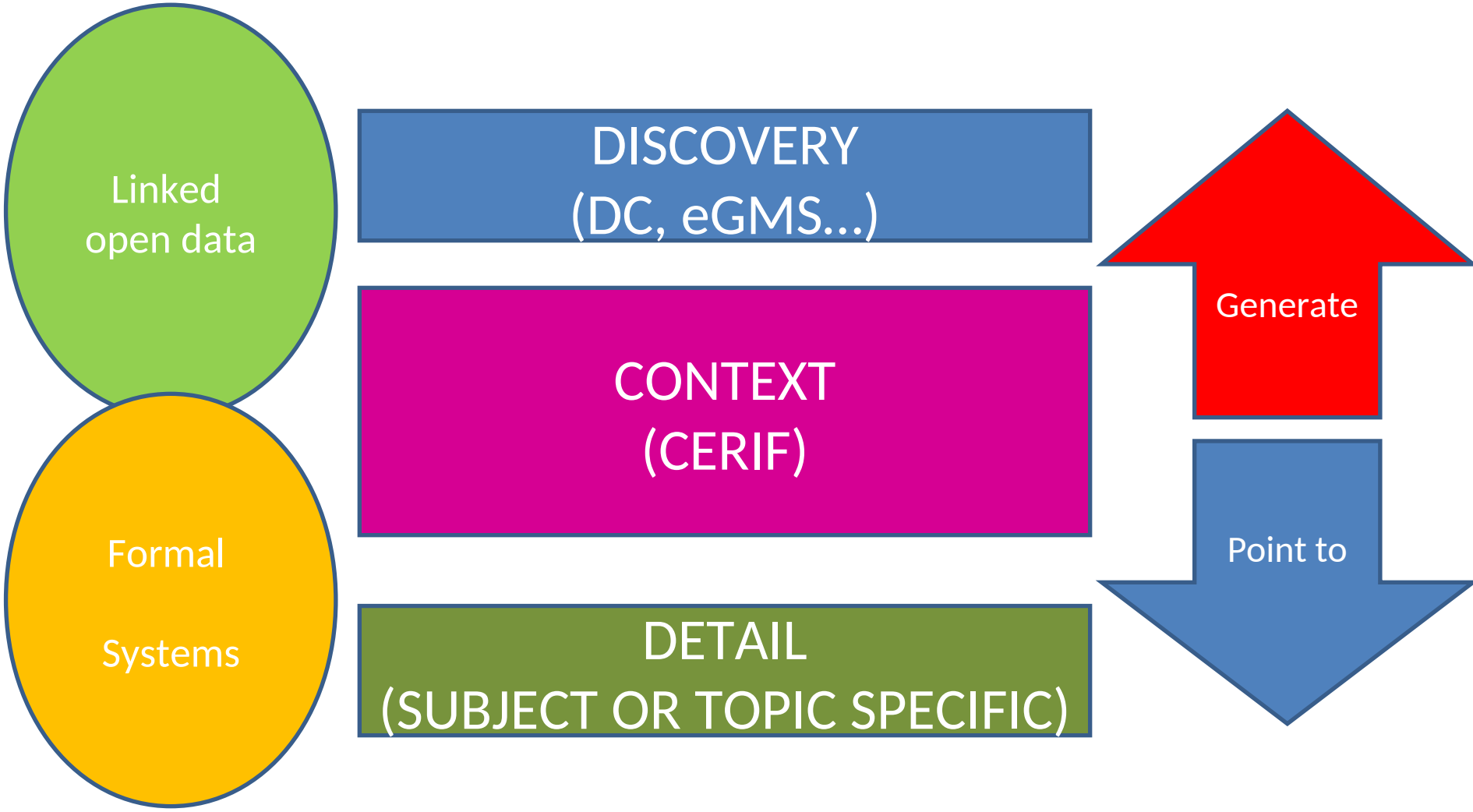
- Most portals clickable lists of datasets
- Most datasets pdf or xls
  - Essentially documents
- Very little metadata
- Metadata 'flat' and poor
- Not linked to underlying research datasets

Unreliable information

**An Opportunity**

**A Problem**

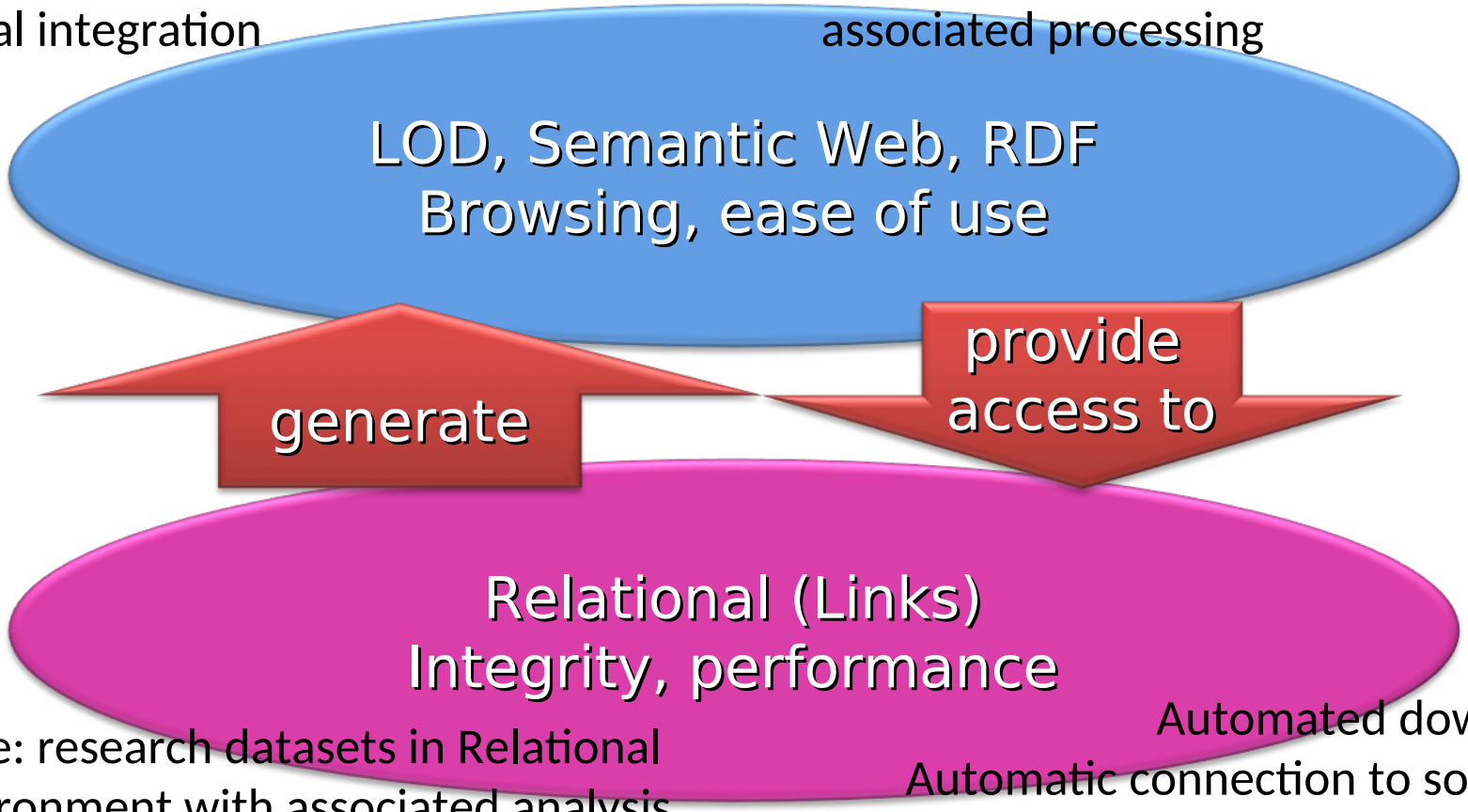
# The Vision: Metadata for Data Model



# processing

Manual download
Manual connection to software
Manual integration

Example: summary data in semantic web/LOD environment (RDF) with associated processing




Example: research datasets in Relational DB environment with associated analysis, visualisation, data mining ….

Automated download
Automatic connection to software
Automated integration

# The Vision: The Models

Complete cohort of researchers, research managers, innovators, media

User Model — interaction with data, processing, persons

Processing Model — providing what the user requires

Data Model — representing research

Resource Model — representing ICT


Complete ICT environment for research

# Conclusion

- Architecture underpinning open data with quality research information
- CERIF provides formality and assurance
- Metadata interconvertors : CERIF superset generating the less rich metadata formats: DC, CKAN…


An Infrastructure for Open, Linked Governmental Data Provision towards Research Communities and Citizens

**The provision of quality metadata assures quality to be confirmed by audit**