

Data Intensive Science: Shades of Grey

Keith G Jeffery^{a *}, Anne Asserson^b

^a Keith G Jeffery Consultants, Shrivenham, SN6 8AH, UK

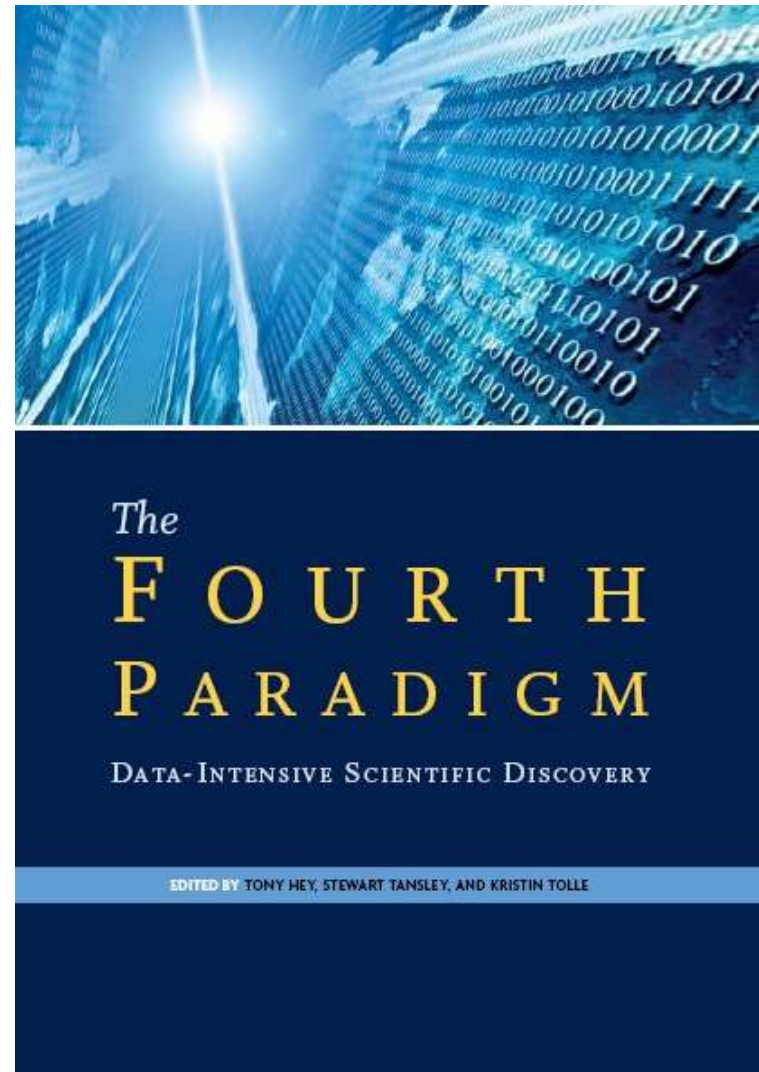
^b University of Bergen, Bergen, 5009, Norway

Structure

- Introduction
- Reliable Information
- Rich Metadata
- Conclusion
- Data Intensive Science
- Grey
- Research Information
- Open Government Data
- Quality
- Context
- Availability
- CERIF
- 3-layer model

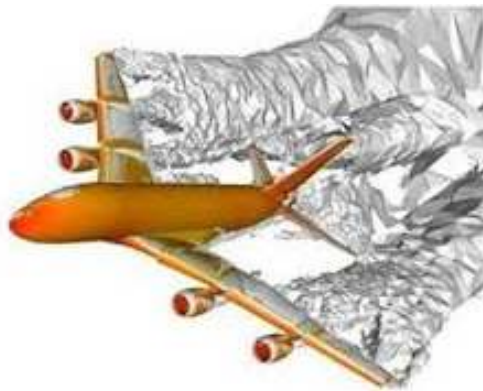
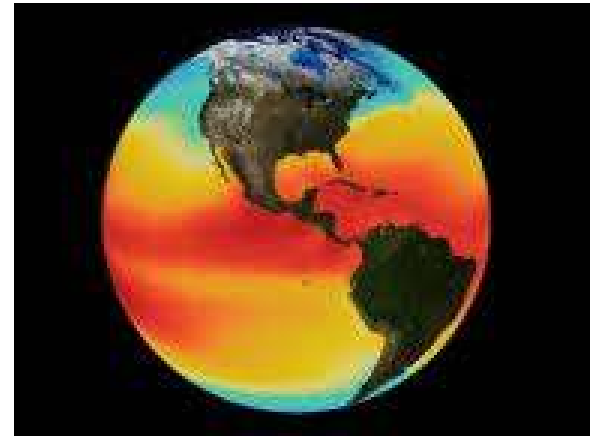
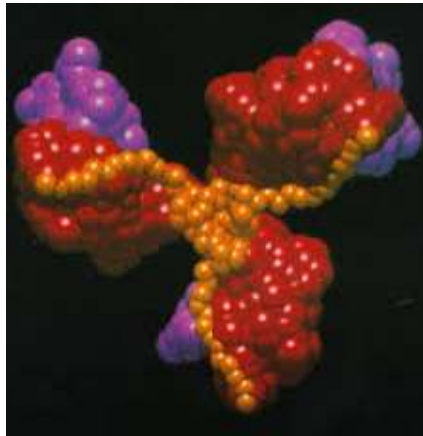
Data Intensive Science

- Research datasets / databases
 - High volume
 - High velocity (of change)
 - Complex structures
 - Streamed
- Data mining
 - Patterns
 - Induction
- Not all ‘patterns’ or ‘rules’ are valid hypotheses



Data Intensive Science

- Related Concepts
 - Open data
 - Available
 - Toll free
 - Big Data
 - Volume
 - Complexity
 - CLOUD Computing
 - Virtualisation
 - Elasticity
 - Pay-as-you-go



Grey

- That which is not white
 - NOT Peer reviewed
- Typically
 - PhD / MS theses
 - Technical reports
 - Lab notebooks
 - Manuals
- But also
 - Newsletters
 - Advertising
- And, importantly
 - Datasets
 - Software
 - Licences
- Patents are peer-reviewed
 - Special process
- PhD theses are peer reviewed
 - Twice if composed of published papers
- Technical Reports undergo internal peer review
 - May be basis of commercial success
- Increasingly research datasets are peer reviewed
 - Especially biomedical

Research Information

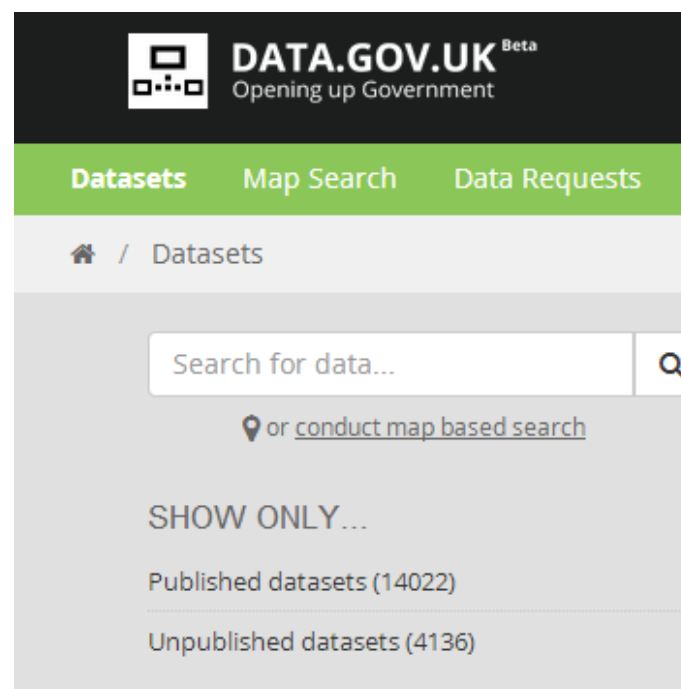
- White Assured
 - by peer review
 - Publishers
 - Impact factor
 - Gold OA Beall's list
<http://scholarlyoa.com/2014/01/02/list-of-predatory-publishers-2014/>
 - San Francisco declaration
<http://am.ascb.org/dora/>



- Grey : how to assure
 - Quality
 - Relevance
 - Access
 - So it can be reviewed
- Review Methods:
 - Usage
 - Citation
 - Annotation
 - Impact (commercial/social take-up)

Open Government Data

- Motivation
 - Transparency
 - Commercialisation
- Derivation
 - Commonly summarised from publicly-funded research
- Vast majority .pdf; then .csv, then .xls
- Metadata DC or CKAN
- ENGAGE Project



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

Structure

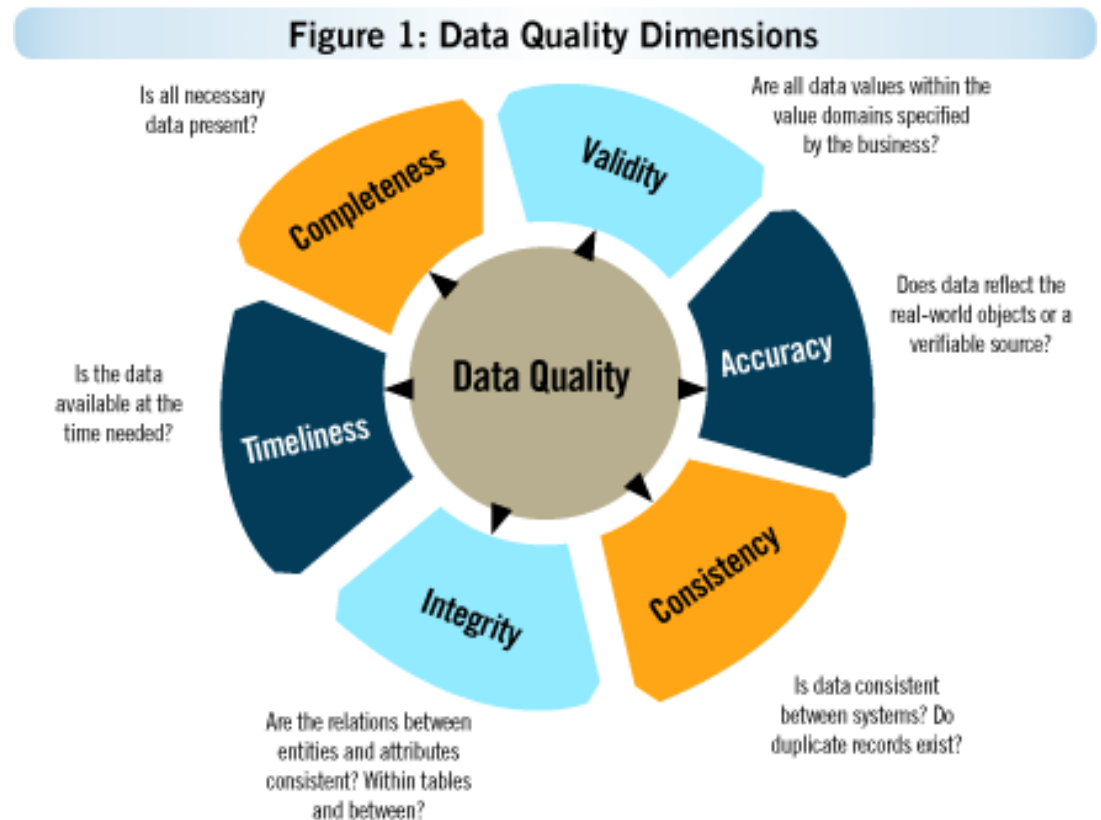
- Introduction
- **Reliable Information**
- Rich Metadata
- Conclusion
- Data Intensive Science
- Grey
- Research Information
- Open Government Data
- Quality
- Context
- Availability
- CERIF
- 3-layer model

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)

Reliable Information: Quality

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



Reliable Information: Context

- Related entities
 - give confidence that the dataset is understood in context
 - Purpose, subject area, research method, associated information
- Used to evaluate dataset for relevance and quality
 - Relevance: Subject area, geospatial / temporal coordinates
 - Quality: organisation, person, publications, facility, equipment, citations

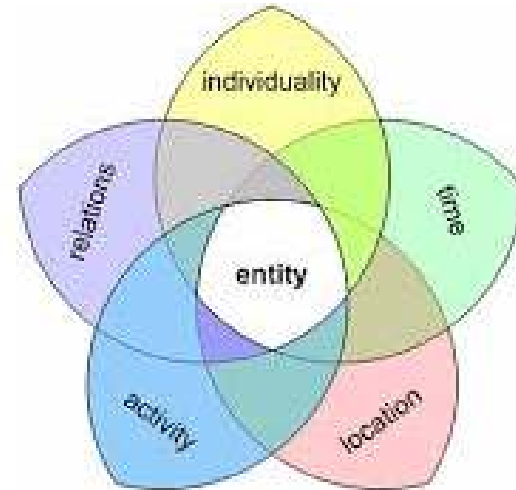


Fig. 1. Five Fundamental Categories for Context Information



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)



Structure

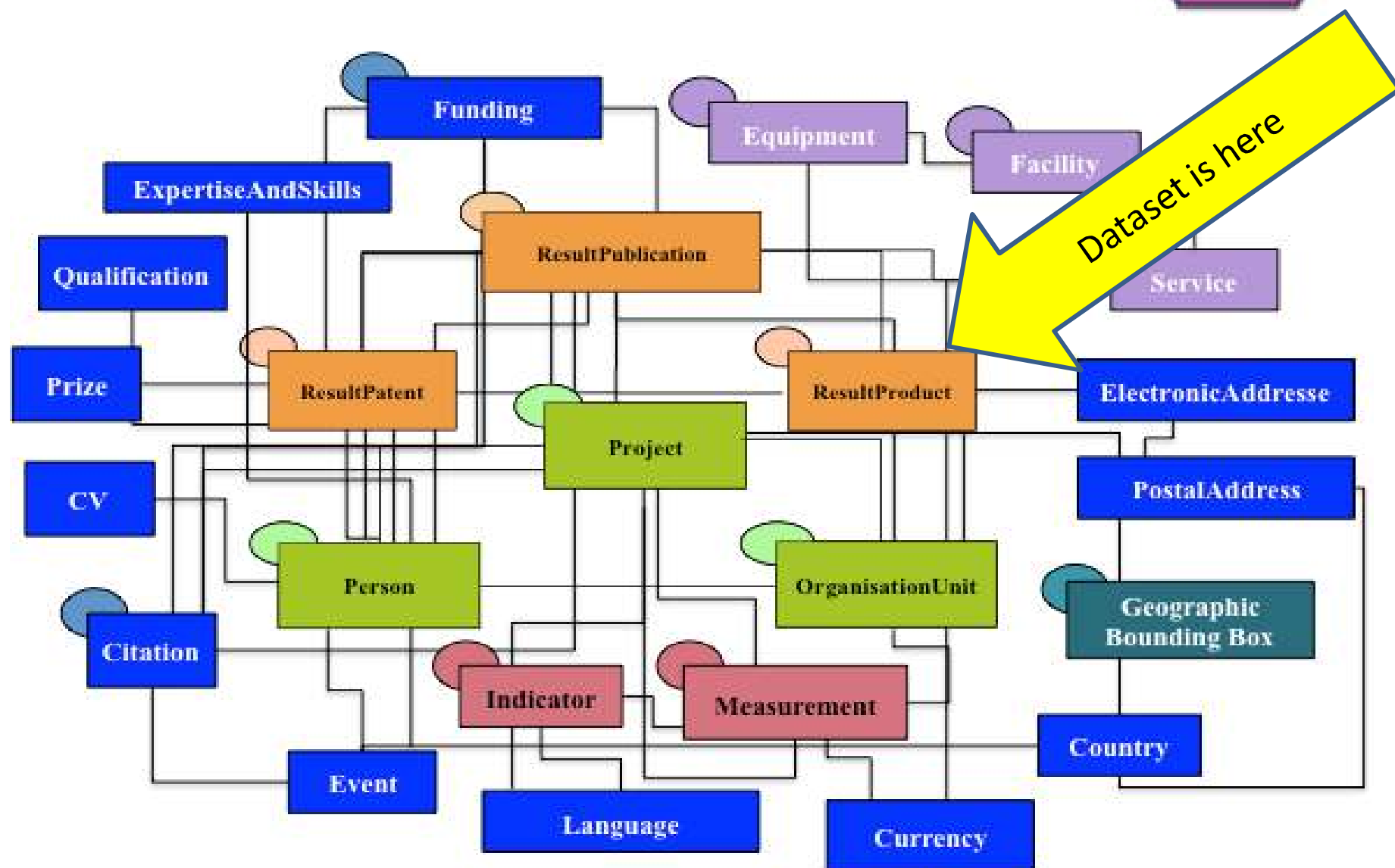
- Introduction
- Reliable Information
- Rich Metadata
- Conclusion
- Data Intensive Science
- Grey
- Research Information
- Open Government Data
- Quality
- Context
- Availability
- CERIF
- 3-layer model

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

CERIF



Dataset Relationships (1)



- Project
- Organisation
 - Collector/creator
 - Owner
 - Funder
 - User
- Person
 - Collector/creator
 - Owner
 - Funder
 - User
- Name, Description, Keywords
- Classification scheme(s)
- GeoBBBox
 - Measurement for precision
- Funding
- Facility
- Equipment

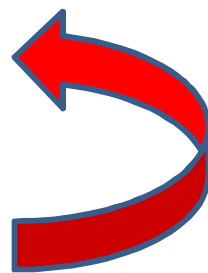
Dataset Relationships (2)



- Publication
 - Scholarly
 - Licence
 - Data Management policy
(including preservation)

- Product
 - Dataset
 - Software
 - Dataset schema

this provides
provenance through
time-stamped role-
based relationships



- Citation
- Measurement
 - Volume
 - Velocity (of change)
 - Accuracy
 - Precision
- Medium
 - classification

Temporal coordinates managed by
linking relations timestamps (e.g.
Project-Product) or if content refers to
an era by classification

Datasets in CERIF: The Debate

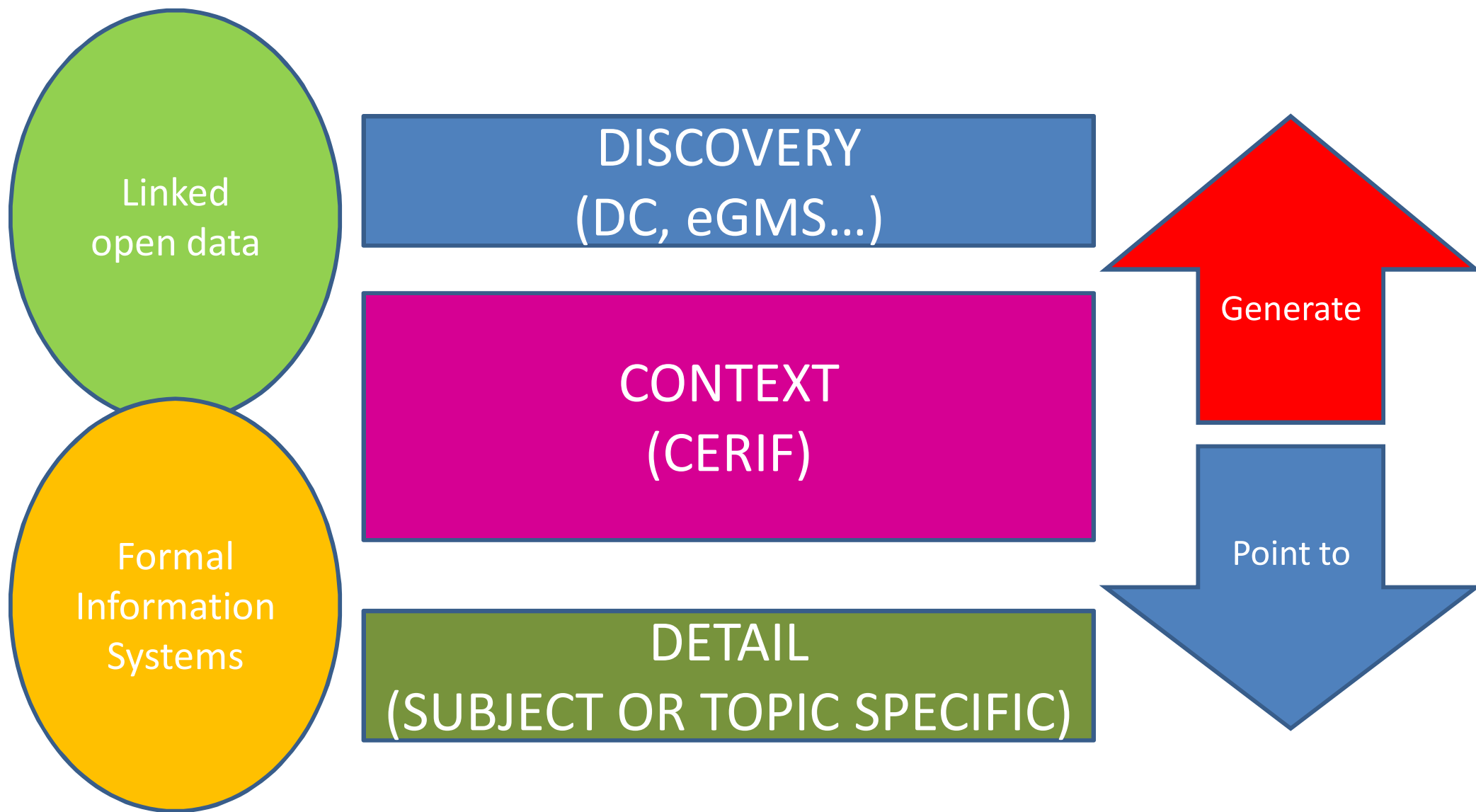
- Keep as Product
 - Seems to work
 - Mainly ‘attributes’ are in linked relations and linking relations
 - If make dataset special what about software, dataset schema,
- Create new entity
 - Gives higher ‘status’
 - Additional attributes required for datasets over product
 - Dataset is important and increasingly so; software not yet

Need:

1. Use cases
2. Mapping to existing CERIF
3. Analyse problems



The Vision: Metadata Stack





engage

Open Data and Information Processing

Manual download

Manual connection to software

Manual integration

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

LOD Semantic Web RDF
Browsing, ease of use

generate

provide
access to

Relational (Links)
Integrity, performance

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

representing research

Data Model

representing ICT

Resource Model

Complete ICT environment for research

We are talking about this

Structure

- Introduction
- Reliable Information
- Rich Metadata
- Conclusion
- Data Intensive Science
- Grey
- Research Information
- Open Government Data
- Quality
- Context
- Availability
- CERIF
- 3-layer model

CONCLUSION

- We assert three points:
 - (1) in the context of data-intensive science the importance of grey;
 - (2) the need for reliability mechanisms to ensure the quality and relevance of grey and
 - (3) the need for rich metadata to support the usage of grey.
- Grey includes research datasets and open government data

USE CERIF FOR DATA INTENSIVE SCIENCE