

< f a s t e r >

BEYOND NESSTAR: FASTER ACCESS TO DATA

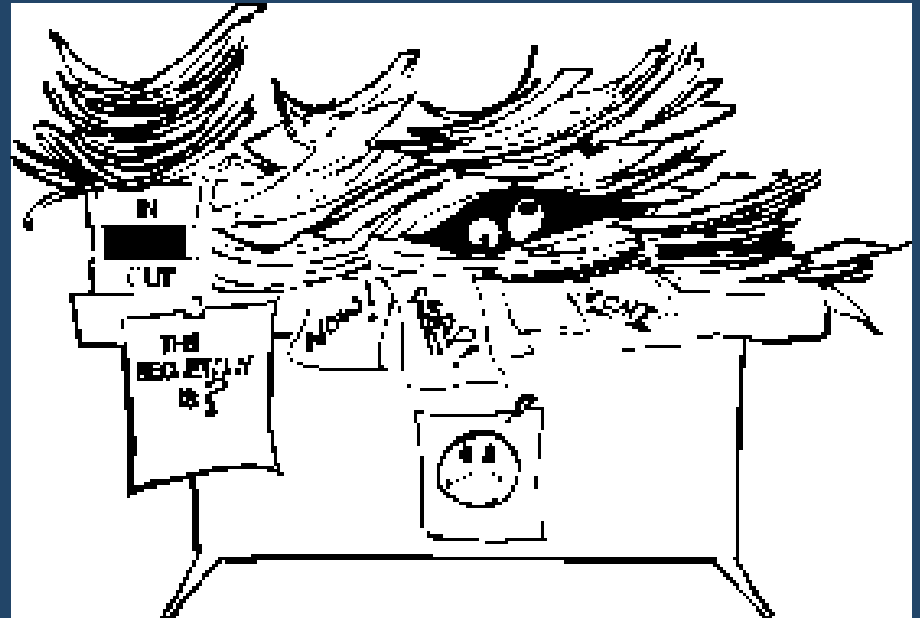
Simon Musgrave, UK Data Archive

Jostein Ryssevik, NSD

w w w . f a s t e r - d a t a . o r g

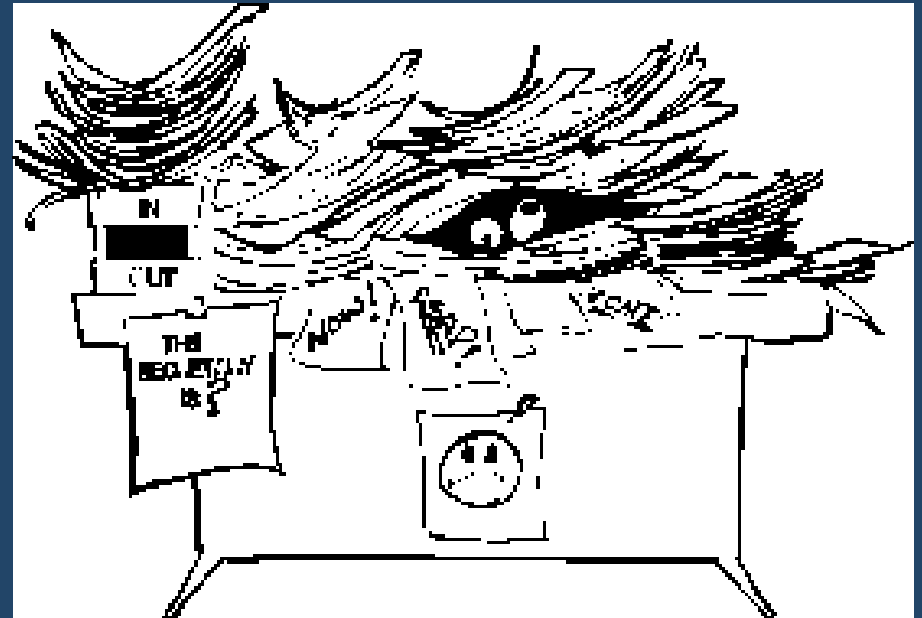
User analysis

- Information overflow - difficulties finding the most relevant information
- Difficulties organising information for efficient retrieval and reuse
- Lack of time
- Lack of time
- **Lack of time** (only a fraction of the day spent on pure research)



So, what are the problems....?

- The system is not **efficient** - it is stealing the professors time
- The system does not **scale**
- The system does not **hyperlink**
- The system does not **publish**

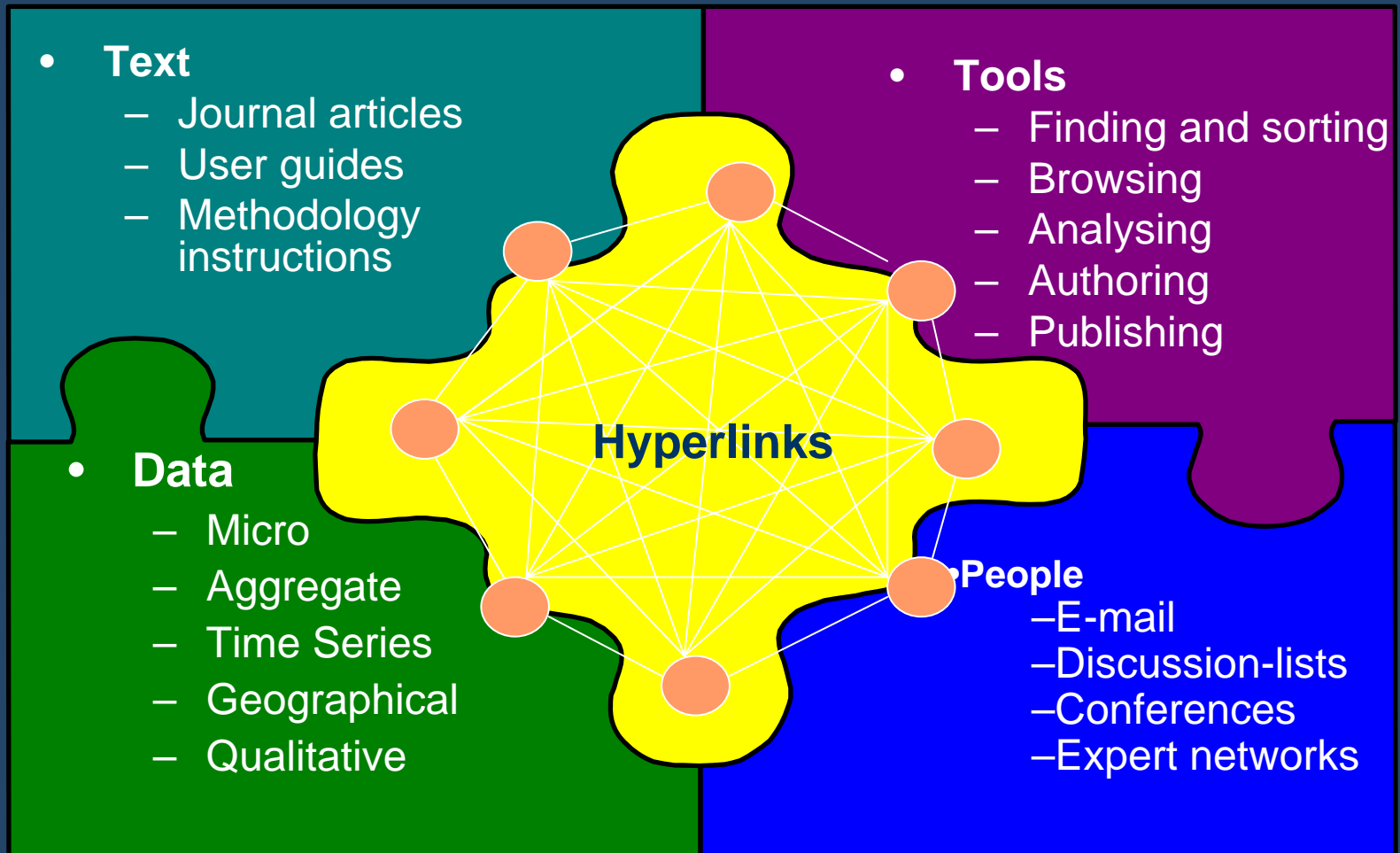


And what is the alternative....?

- A system so **efficient** that it can release time for productive work
- A system that **scales** - is able to handle any amount of information
- A system that allows efficient **hyperlinking**
- A system that is open and allows **publishing**



...a social science Workbench...?



“The Dream List”

- all existing empirical data available on-line
- an integrated resource discovery gateway that could help to identify and locate these resources
- extensive amounts of metadata available (multimedia, hyperlinked and totally integrated with the data)
- the ability to browse, analyse and visualise data on-line
- the ability to convert the data in one of a number of formats and copy, with the metadata, to a local machine
- "active research agents" (knowbots) mining the net and informing the user when new data within their special field of interest are made available
- efficient hyperlinks from the data sources to every scientific publication ever produced on the basis of a dataset
- ditto e-mail/web addresses to all relevant researchers, departments etc.
- an efficient feedback system to the body of metadata allowing the user to add to the collective memory of a dataset

NESSTAR 1.0 features

- An architecture for a totally distributed virtual data library
- The ability to locate multiple data sources across national boundaries
- The ability to browse detailed information about these data sources
- ..and to do simple data analysis and visualisation over the net
- ..or to download the appropriate subset of data in one of a number of formats
- Allowing the user to bookmark resources in the data and metadata repositories
 - searches
 - datasets
 - analysis (tables, models etc.)
- ..and to hyperlink these resources from external Web-objects (like texts)
- ..or to “subscribe to” bookmarks and leave them with the “digital research assistant” for automatic and regular execution
- A system for imposing a variety of access control policies
- A system for remote publishing of data to NESSTAR servers

FASTER Objectives

(not only NESSTAR at a higher speed)

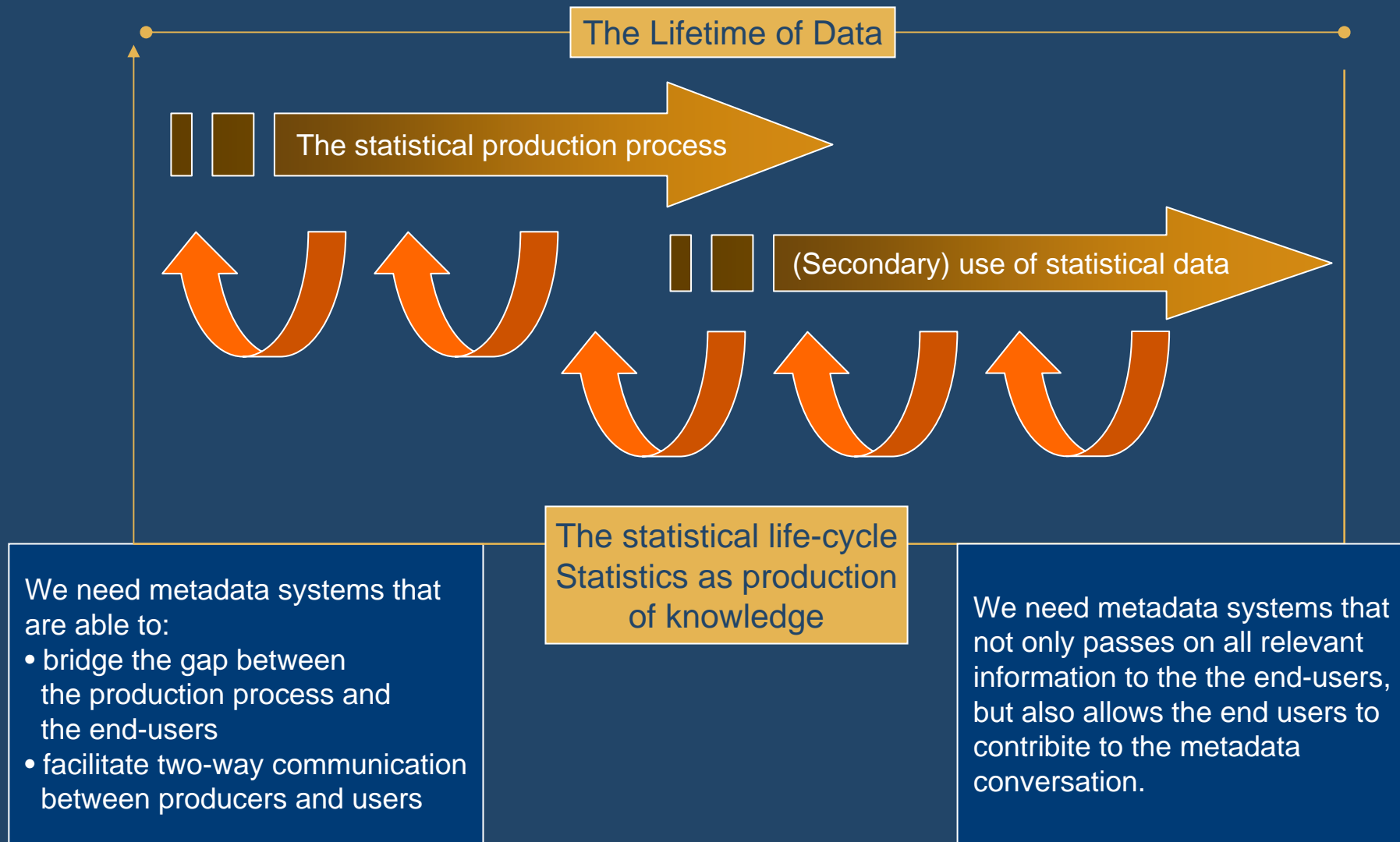
- To move beyond rectangular survey data
 - hierarchical data, time-series data, multidimensional tables etc
- To strengthen the link to the statistical production process
- To provide a flexible and extensible metadata object model to support the functionality system
- To develop a highly flexible client interface that is able to respond to needs of the user as well as the needs of the data (as described by the metadata)
- Develop a secure access control system that is driven by the metadata
- Integrate (automated) statistical disclosure control driven by the metadata

Metadata

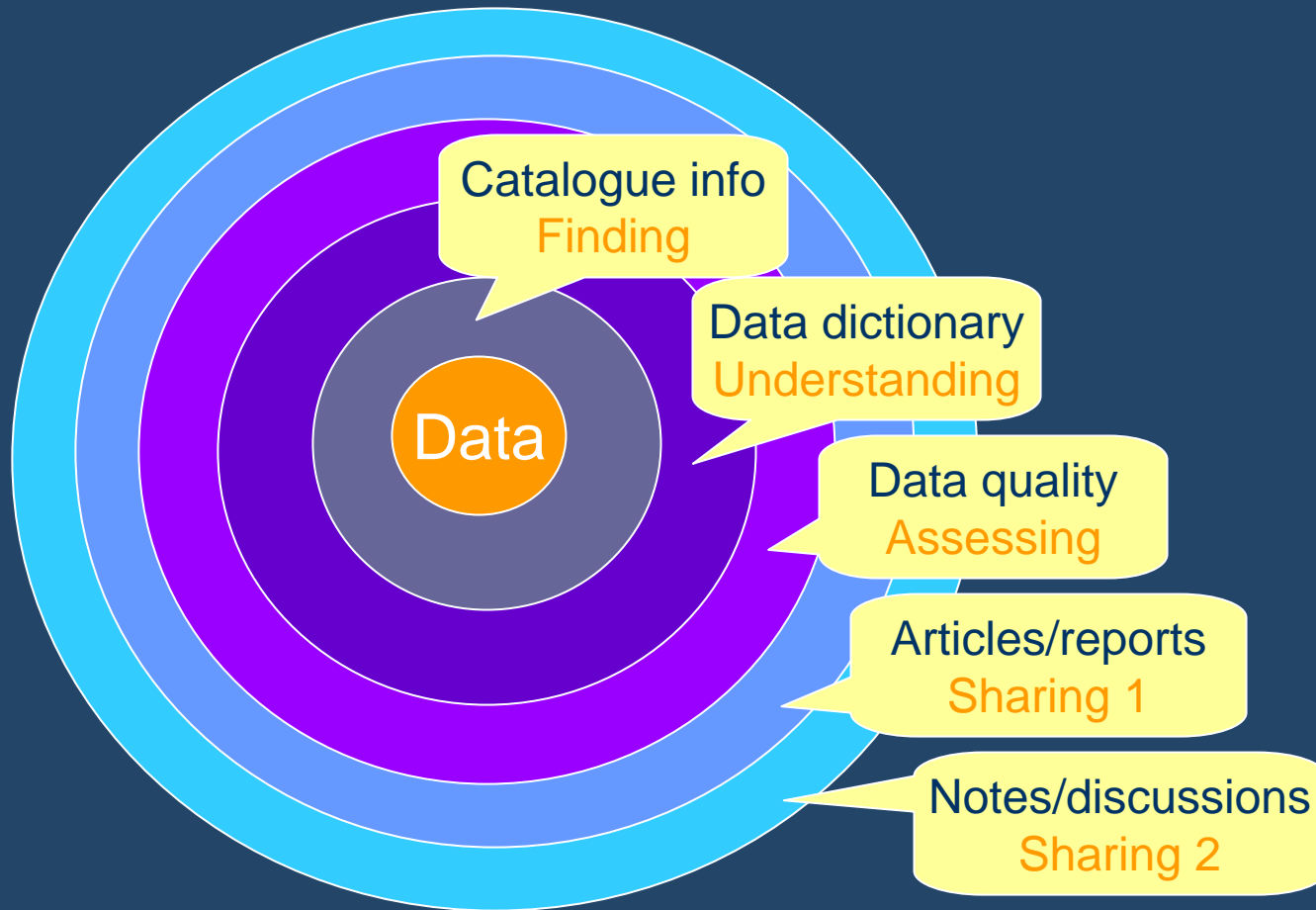
- the glue of the Data Web

- Machine understandable metadata
 - providing knowledge about data to software processes (configuring interfaces, driving transformations, sub-setting, access control, disclosure control etc.)
- Human understandable metadata
 - **Finding**: metadata used for resource discovery
 - **Understanding**: metadata used to inform the user about the content and meaning of data/numbers
 - **Assessing**: metadata used to inform about the quality and limitations of a data source
 - **Sharing**: metadata as a conversation between persons offices and organisations working with a dataset

Metadata as communication



...layers of metadata

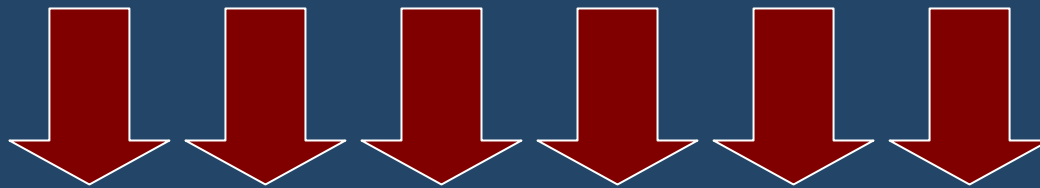


..a note on the dependencies between NESSTAR/FASTER and the DDI

- Given the ambitions of the NESSTAR project it is difficult to see how the system could have been built without a widely accepted, highly structured and over-all Web-friendly metadata standard like the DDI.
- Without agreement on metadata standards, we can't even start talking about interoperability across information systems
- The DDI-standard is also dependent upon software systems (like NESSTAR) to prove its usefulness. Without productivity gains or improvements in the quality of products or services, acceptance of a new standard might be hard to justify. This can only be achieved through software support.
- Software implementation is also an excellent way to detect shortcomings or ambiguities in a standard.
- A parallel: Without Tim Berners-Lee's first version of HTML, Marc Andreessen and his group at NCSA would not have made a fortune on developing Mosaic and eventually Netscape. Likewise, without the development of the first Web-browsers (like Mosaic and Netscape), HTML would probably have remained a local hypertext dialect for technical documentation at Cern, Switzerland

Adaptable interface

Types of users/user preferences



Types of data/resource

Access control

The basic dilemma:

How to make an access control system that combines easy access to data without violating legitimate needs to protect data against unauthorised use

or

.....how to please the data owners without frustrating the users.

Two layers of control

- **Access control:** A decision mechanism that decides whether or not a particular user has the right to perform an operation on a specific data source or not.
- **Statistical disclosure control:** Automated suppression of detailed information to avoid disclosure of confidential information from micro-data.

Access control in world of mobile data and global users

What decision should the access control system make when a Norwegian researcher staying temporarily at a Canadian university requests to download a Danish dataset stored at UK Data Archive?

The importance of making it right...

An access control system that is too liberal (too many “yes”) will frustrate the data depositors.

An access control system that is too restrictive (too many “no”) will frustrate the end users.

In both cases the legitimacy of the archives/libraries will be undermined!

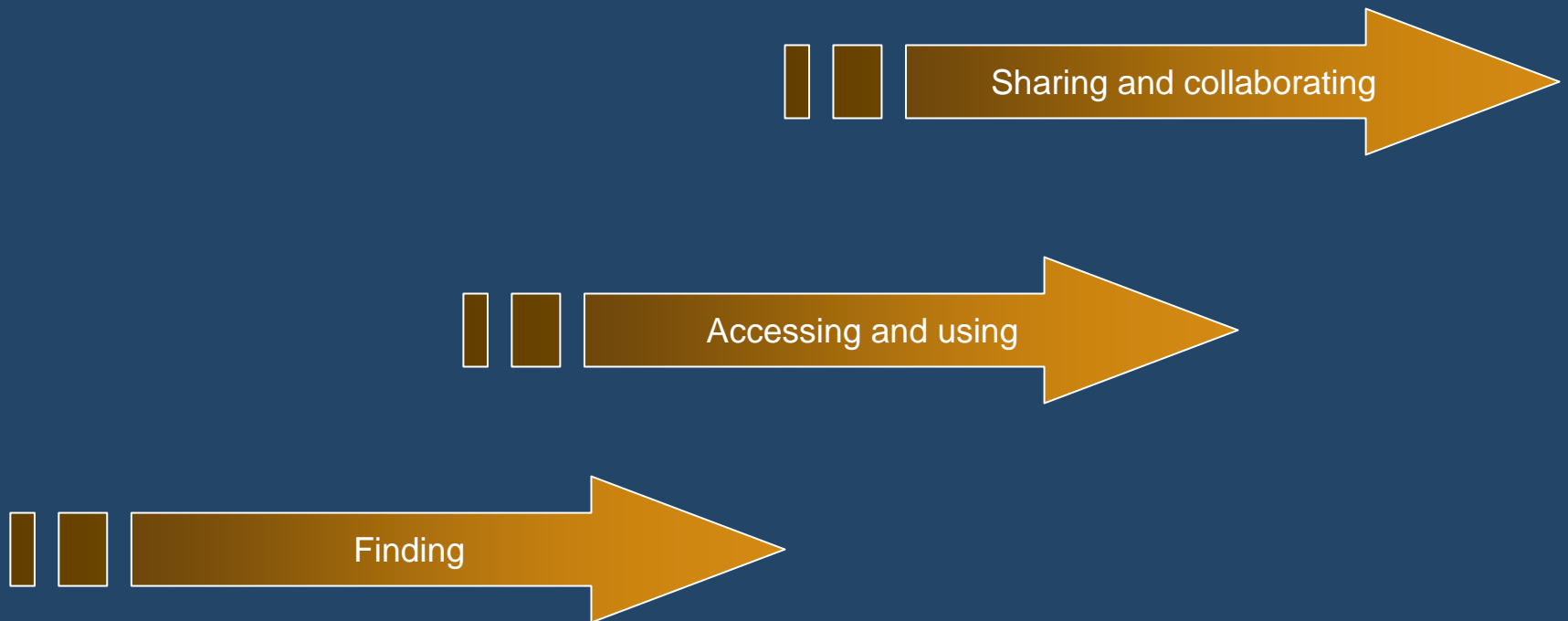
FASTER partners

- UK Data Archive
- Norwegian Social Science Data Services (NSD)
- Danish Data Archive (DDA)
- Statistics Netherland (CBS)
- University of Milano (Dipartimento di Scienze dell'Informazione)
- Central Statistical Office (CSO) Ireland
- Statistics Norway (SSB)
- Centre d'Informatisation des Données Socio-Politiques (CIDSP), France

Complementary developments

- **LIMBER** - Language Independent Metadata Browsing of European Resources
 - development of a multilingual social science thesaurus to guide searches in multilingual metadata repositories
 - ...and to support semi-automated indexing of metadata elements

Data Archives - from data graveyards to data greenhouses



Target Users

- Researchers
- Students
- Librarians
- Policy Makers
- Journalists



all want access to their personal(ised) workbench

Metadata Content (3)

- **Contextual (DDI level 5 +)**
 - the sky's the limit
 - background - user guides, questionnaires
 - teaching and learning
 - concepts
 - multi-media descriptions (e



Metadata Content (4)

- **Quality (DDI ?)**

e.g.

- Methodology
- Response rates
- Provenance
- Processing procedures

Metadata Content (6)

- **Bookmarks/hyperlinks**
 - searches
 - datasets
 - analysis (tables, models etc.)
 - download

Typically material added by users of the data