From Data Graveyards to Knowledge Greenhouses

Simon Musgrave (UKDA)
Jostein Ryssevik (NSD)
What we have... is a lot of data
But maybe it is...

....a graveyard
What we need... is a Knowledge Greenhouse
...to get something like this
Introduction

- The data graveyard
- Keeping data alive … realising their full potential
- The knowledge greenhouse
- Theory and practice, the development of the NESSTAR dreams
Maslow's basic position is that as one becomes more self-actualized and transcendent, one becomes more wise (develops wisdom) and automatically knows what to do in a wide variety of situations.

James (1892-1962) hypothesized the levels more simply as material (physiological, safety), social (belongingness, esteem), and spiritual

- "Where there is no vision, the people perish." Proverbs 29:18
- "Ah, but a man's reach should exceed his grasp, Or what's a heaven for" 'Andrea del Sarto' by Robert Browning
MASLOW'S Hierarchy of Needs

- Physiological Needs
- Safety Needs
- Belongingness & Love Needs
- Esteem Needs
- Need to Know & Understand
- Aesthetic Needs
- Self-Actualization
- Transcendence
Need hierarchy and data hierarchy

Human needs (James)
- Spiritual
- Social
- Material

Data needs
- Knowledge elicitation
- Interaction
- Preservation
Preservation - material

- Role of archives to preserve data
- Environmental conditions (e.g. BS7799 standards for information security management)
- Software and system independence
- Safety from external attack and internal error
- On-going management and migration
- Data dies from neglect
Simply preserve?

- We know that preserved data are not dead - but are they fulfilling their potential?
- Survival is not the limit of our vision
- Data die through loneliness, they are social
- The more they get used, the less they decay and the more valuable they become
Data Interaction - social

- Developments to support interoperability
  - Structure (DDI, Dublin Core, RDF))
  - Semantics (CESSDA group, LIMBER)
  - Syntax (XML)
- XML and DDI are well-established, the semantics may be the biggest challenge
- Data can be seamlessly embedded in a variety of objects
Political Environment

- Data thrives in a distributed not centrally controlled environment
- Data are best supported and released by those who love them and know them best, the data owners or distributors – keep Norwegian data in Norway
- The risk is higher but, like people, data thrive via delegated structures and agreed standards of behaviour
Knowledge Elicitation

- New human needs (cognitive, aesthetic, self-actualization, transcendent) drive our vision and demand for knowledge to enhance our wisdom.
- Simultaneously (and co-determinantly) new technologies emerge to enable our vision to become a reality.
The knowledge greenhouse

- To elicit knowledge we need to create the right environment
  - Care and attention (management and migration)
  - Freedom from disease (bugs and errors)
  - Conditions for growth - fertiliser, heat, light, water (the right interoperable environment)

- We need to be able to add value and link complementary resources
  - Pedagogical material
  - Contextual information
  - Scientific framework
  - Social and economic environment
...data in the knowledge production process

The Lifetime of Data

The statistical production process

(Secondary) use of statistical data

Knowledge
it’s all about communication
User scenarios from the Knowledge Greenhouse

- A user analysing a group of variables in dataset X would like to know if there are similar datasets from other countries that could be used for a comparative study.
- She would also like to have an overview of knowledge products (papers, articles etc.) based on this study and even to browse these objects if they are available online.
- Moreover, she would like to contact other researchers that have used the dataset to hear about their experiences.
- Finding a problem with one of the variables, she writes a note and appends it to the “user experience-section” of the metadata to alert future users (she also leaves her e-mail address to allow them to contact her).
- And when the research paper is ready and published in an online journal, links to the dataset is added to allow future users to revisit her analysis.
...more scenarios from the Knowledge Greenhouse

- A user that is reading an article in an on-line journal finds a link that connects him to the data that was used by the author to underpin the argument. The link allows the user to rerun the analysis, and also to dig deeper into the same data-source.
- He is also made aware of several other data sources published after the article was written and he uses these to challenge the conclusion of the author.
- Links to knowledge products based on these newer data sources is also available.
- From one of the sources he is even brought to a mail-list that discusses the phenomena in further detail.
...even more scenarios from the Knowledge Greenhouse

- ...a user is looking at a table showing variation in nationalistic attitudes among different educational groups in Norway.
- ...through a multilingual thesaurus service he is able to pick up the relevant key-words describing this table and to automatically create a multilingual query for datasets that might be used to create comparable tables.
- ...he also leaves the query with his "digital research assistant" (an active agent), to make sure that he is alerted if a new dataset meeting his requirements is published somewhere around the world at a later stage.
- ...he even asks his agent to look for other digital objects addressing the same topics.
The Web dream comes true.....

- The current Web technology is taking us a long way towards the realisation of these dreams
  - From “one to many” to “many to many”
  - From publishing to collaboration
  - From many local to a single global hypertext-space
  - The Web has taken all existing media as its content (real multi-media)
  - The Web has memory
  - The Web has the “right” amount of standardisation
...but still some missing bits and pieces

- The Web is still poor on semantics. Most of the resources on the Web is ment for human consumption.
- The natural next step in the development is the "semantic Web", the Web that allowes us to describe digital resources in such a way that the resources can start talking to each other and to software processes.
- "The Data Web" that we are dreaming about is the statistics department of this general Semantic Web.
...the DDI 1.0

- The biggest achievement of the data archive World
  - **Acceptance:** fast take-up in the community of data archives and data libraries world-wide
  - **Community building:** revitalised the co-operation and sharing of know-how and technologies among the archives and libraries
  - **Strengthening of the ties to the data producers**
  - **Software development**
...beyond DDI 1.0

.... still some challenges

- **A pure “bottom-up” approach:** The DDI is used to describe concrete files or products coming out of the statistical process. It has no level of abstraction above or beyond a physical statistical product.

- **The “study” (survey-instance) as the highest level:** There is no way to describe relationships between data elements/variables across studies.

- **Extensibility:** The DTD is a non-extensible construction, if you need to make an addition you either create a new one or you break it.

- **Machine-understandable versus human-understandable:** Using XML does not automatically create metadata that is complete and logical enough to drive software processes.
...elements of the Data Web (the foundation of the Knowledge Greenhouse)

- DDI 2.0... the more modular, extensible and machine-understandable version of the DDI
- Domain specific ontologies, thesauri and controlled vocabularies that will allow us to add machine-understandable and Web-accessable semantics to our DDI-described data
- ...expressed in an standard framework like RDF that will allow us to create mappings between domain specific ontologies
- Software systems that are able to handle this semantics
- ...and a lot of hard work to mark-up and describe our existing resources
...so where are we and where are we heading?

- DDI 1.0 is here and is taken up quite rapidly in the community
- ...and the DDI 2.0 process is in the pipeline
- ...a social science multilingual thesaurus is being developed within the LIMBER project to allow intelligent language independent classification and searching of social science resources.
- ...the LIMBER thesaurus will interoperate with DDI metadata (adding semantics and controlled vocabularies to the metadata)
- and is expressed in RDF to allow easy mappings to other domain specific thesauri
....and

- Software systems are developed or under development to make resources described by the standards come to live
- NESSTAR 1.1 is already here and used to run live data services in a few European data archives.
  - 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
...and

- 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 remote publishing of data to NESSTAR servers
- ..a Web engine that allows user to access NESSTAR resources through a standard Web-browser

The NESSTAR technology is further developed within the FASTER project that among other things will add integrated support for tabular/aggregated data
...NESSTAR not the only system...

- there is a lot of Knowledge Greenhouse building going on out there..
  - ILSES
  - FERRET (US Census)
  - Virtual Data Library (Harvard)
  - WebDAILS

- The important thing is that we are basing our systems and resources on the emerging open standards so that we can allow systems as well as data to talk to each other.
..then we can all meet and have fun in the.....
..however

..as we know that the road from the Data Graveyard to the Knowledge Greenhouse is paved with a lot of hard work and sleepless nights, we would like to end this session by playing you a blues....

...a metadata blues