

Threats and outline of the Roadmap

David Giaretta, STFC

Motivation

- What is a roadmap?
 - What is needed to support wide range of uses
 - “How to do we get from where we are to where we want to be?”
 - Directs development of the parts of the e-infrastructure needed to assure the continued usability and accessibility of scientific information

Objectives

- The original view of roadmap development
 - Review and synthesize existing roadmaps

Inputs to draft Roadmap

- Alliance Research programme
- Alliance action plan
- CASPAR Conceptual Model
- JISC Digital Repositories Roadmap
- DCC Lifecycle model
- DPE Research Roadmap
- Warwick workshop report: Digital Curation and Preservation: Defining the research agenda for the next decade
- Requirements for Digital Preservation Systems: A Bottom-Up Approach (2005)
- Thirteen Ways of Looking at...Digital Preservation (2004)
- Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years (2002)
- Digital Preservation and Deep Infrastructure (2002)
- Report of the Task Force on Archiving of Digital Information (1996)
- Mind the Gap report from the UK DPC
- E-INFRASTRUCTURE STRATEGY FOR RESEARCH: FINAL REPORT FROM THE OSI PRESERVATION AND CURATION WORKING GROUP
- eIRG Roadmap
- ESFRI Roadmap
- Developing World-class Research Infrastructures for the European Research Area (ERA) - report of the expert group
- CASPAR test case questionnaire
- Cyberinfrastructure Vision for 21st Century Discovery
- Invest to Save
- It's About Time
- Stewardship of Digital Research Data: A Framework of Principles and Guidelines
- To Stand the Test of Time - Long-term Stewardship of Digital Data Sets in Science and Engineering

..Objectives

- The original view of roadmap development
 - Review and synthesize existing roadmaps
 - Produce draft on this basis
 - Refine in the light of evidence
 - Used to provide some structure to Surveys
 - BUT needed to be careful to avoid pre-determining the responses

Threats

- Use threats/risks
 - Allowed us to structure part of the Survey
 - Leads “easily” to technical solutions
 - Strawman to argue about
 - We need to be flexible to allow new threats/risks to be added

....Objectives

- The revised approach
 - Excellent response to surveys allowed evidence to be incorporated earlier
 - The centrality of threats: link between disciplines and roadmap
 - Users may be unable to understand or use the data e.g. the semantics, format, processes or algorithms involved
 - Non-maintainability of essential hardware, software or support environment may make the information inaccessible
 - ...
- D2.1 “Draft roadmap” delivered

Roadmap outline

- Components of roadmap
 - Financial
 - Organisational/social
 - Policy
 - Technical
 - (plus virtualisation as underlying concept)

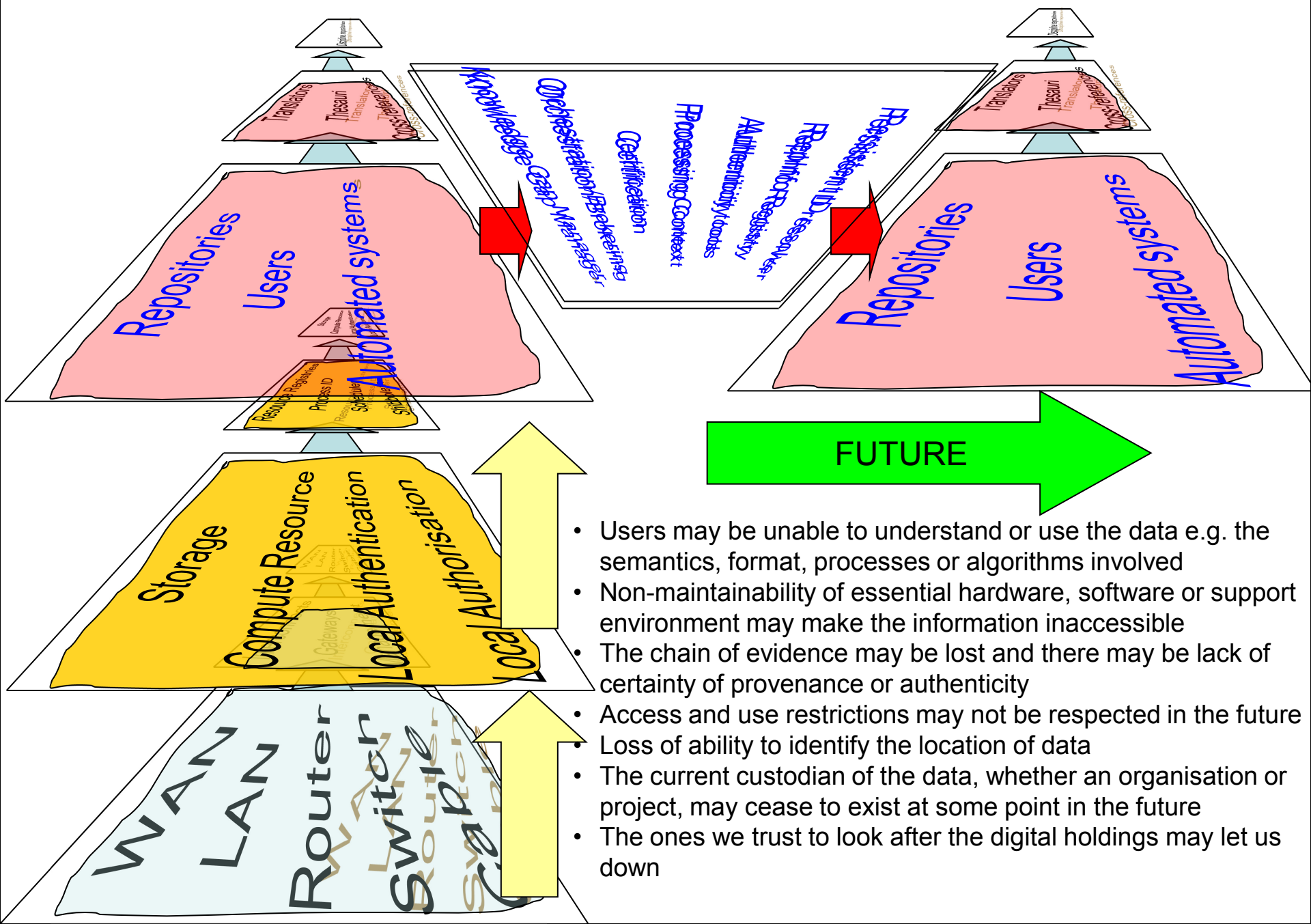
- Phases:
 - Prototype
 - Early adopters
 - Stable
- Business Model??
- Scaling
 - Some aspects grow at least as fast as the amount of data
 - Other components may grow more slowly
- Cost models
- Benefit models

Organisational and Social

- Responsibility?
- Fears of misuse or legal issues
 - Possibly unfounded

- Deployment and adoption
- Bottom-up or top down?
- Certification

Threat	Requirement for solution
Users may be unable to understand or use the data e.g. the semantics, format, processes or algorithms involved	
Non-maintainability of essential hardware, software or support environment may make the information inaccessible	
The chain of evidence may be lost and there may be lack of certainty of provenance or authenticity	
Access and use restrictions may make it difficult to reuse data, or alternatively may not be respected in future	
Loss of ability to identify the location of data	
The current custodian of the data, whether an organisation or project, may cease to exist at some point in the future	
The ones we trust to look after the digital holdings may let us down	



Technical aspects structure

- Scenario
- Next steps
- Final destination
- Relevant projects. Policies, organisations, activities

- A dataset created by one researcher may need to be used by a second, either contemporaneously or at some later time. This second researcher may come from a different discipline and use different analysis tools. In order to avoid producing misleading results he/she must be able to understand what the data actually means. For example, given an astronomical image in the current FITS format, with its several variants, the researcher would need to be able to extract the values of the pixels of the image from what may be quite a complex and highly tailored digital object. In order to use an analysis tool one would need to know how to deal with these pixel values, their units, their coordinates on the sky and the way in which the photons have been selected e.g. the bandpass of the filters used. Representation Information is the OAIS term for everything that is needed in order to understand a digital object. A registry would help to ensure that the required Representation Information is available in the future and across disciplines.*

Next Steps

- Representation Information Registry holding copies of Representation Information of all types which can be shared and enhanced by contributions from many people.
- Virtualisation techniques to facilitate easier integration into contemporary tools
- Preservation features should be embedded in the "creation" environment, automating/facilitating the generation of necessary representation information (data, models, assumptions, configurations, ...).
- Knowledge Gap Manager which provides a semi-automated way of identifying where additional Representation Information needs to be created, based on information collected by the Orchestrator/Broker
- Processing Context which helps to maintain information about the processing history of a dataset

Relevant projects, policies, organisations, activities:

- **CASPAR**
(<http://www.casparpreserves.eu/>) ,
- JISC (<http://www.jisc.ac.uk/>),
- OAIS
(<http://public.ccsds.org/publications/archive/650x0b1.pdf>),
- SHAMAN (<http://www.shaman-ip.eu/>),
nestor
(<http://www.langzeitarchivierung.de/>)

Final destination

- A set of services, supported over the long term, which make it easier to maintain adequate Representation Information, particularly after active work on the dataset has ceased or slowed. Automated capturing of the creation and processing context.

Threat	Possible Solutions
Users may be unable to understand or use the data e.g. the semantics, format, processes or algorithms involved	Representation Information Toolkits <input type="checkbox"/> Descriptions down to the bit level & Semantics <input type="checkbox"/> Representation Info registries
Non-maintainability of essential hardware, software or support environment may make the information inaccessible	Broker of information Emulation capabilities Virtualisation and Access Software
The chain of evidence may be lost and there may be lack of certainty of provenance or authenticity	Authenticity Model and tools
Access and use restrictions may make it difficult to reuse data, or alternatively may not be respected in future	Digital Rights and Access Rights virtualisation
Loss of ability to identify the location of data	An ID resolver which is really persistent
The current custodian of the data, whether an organisation or project, may cease to exist at some point in the future	Brokering of organisations to hold data and the ability to package together the information needed to transfer information between organisations ready for long term preservation
The ones we trust to look after the digital holdings may let us down	Audit and Certification based on RAC document (currently undergoing ISO review)