Digital Provenance

Roger Layton

The ETHER Initiative,

Johannesburg, South Africa

OpenCulture 2012, Kia Oval, London : 26 June 2012. Roger Layton "Digital Provenance".

roger.layton@ether.co.za

ETHER

What is Provenance?

- Tracing the history back to the maker (Provenance as HISTORY)
- Continuous and unbroken line of valid evidence (Provenance as EVIDENCE)
- Physical Provenance – as a sequence of verbs (Provenance as ACTIONS)
  - Created
  - Sold
  - Exhibited
  - Documented / Analysed
  - Evaluated
  - Loaned
  - Stored

Objects of Provenance

- Tangible heritage : museum objects, buildings, sites
- Documentary heritage : content within the objects (records, letters, diaries, notes, journals, books, photographs, video, audio, …)
- Intangible heritage : oral history, indigenous knowledge, rituals, dance, performances, …)
- Digital heritage : databases, digital records, digital documents, emails, web sites, blogs, …

roger.layton@ether.co.za

ETHER

Why is Provenance important?

- To establish authenticity and trust
- To establish the basis for rights
- To communicate the full history
- To distinguish fraudulent copies vs originals
- To ensure that future generations will be able to trust the digital heritage we leave as our legacy to them

Digital Heritage – Digital Provenance

- All digital heritage is essentially digital documents / files in a folder on some device (hard drive, memory stick, DVD, cloud)
- Can be easily copied if not protected
- All digital provenance information is also a digital document
- This can also be easily copied and modified
- QUESTION: How to ensure authenticity of the provenance information?

roger.layton@ether.co.za

ETHER

Challenge of provenance

- All heritage is becoming digital – in 20-50 years this will be the ONLY history that people know
- The world's storage is increasing by factor 1000 every 10 years
- We can storage anything any number of times – but which is the definitive version, and how long will this persist?
- How to trace the history which include digital objects?
- New standards and practices for digital heritage are required to complement SPECTRUM
  - Digital Heritage Body of Knowledge (DHBOK) – as a 10-step process for handling of digital content

roger.layton@ether.co.za

ETHER

Challenges in our work

- In creating digital repositories – we need to address digital provenance for the long-term future
- Our focus has shifted
  - From digitisation strategy and techniques
  - To exploring how to best document digital provenance
- A sample of 3 projects are outlined…

roger.layton@ether.co.za

ETHER

Project 1 : ǂKhomani San

- The San (bushmen) are the **oldest tribe in the world**
- They will be **lost as living heritage** within 20 years
- for the past few years various trips have been made to collect material from the remaining bushman
- **Oupa Dawid**, the elected representative of the Khomani San, **died on 13 June 2012** one of the last speakers of the language and of the culture of these "first people"
- Goal of project : to help to **inventorise and digitise the materials collected from various sources** and in various institutions and various forms and formats – to create a virtual digital collection

Khomani San : Challenges

- The bushman people have been in Southern Africa for the past 20,000 years at least + traceable back to cradle of humankind in Krugersdorp
- Can we build a digital archive which will last 20,000 years?
- **Provenance Challenge:** How to build provenance into the digital archive as it moves through many future digital migrations

Project 2 : Khulumani

- Khulumani : a social movement for the human rights of survivors of apartheid-era gross human rights violations
- A large oral history archive (> 30,000 stories)
- Our role : create a massively-connected semantic digital archive
- **Provenance challenge :** long-term authenticity of the oral history records – traceability back to original collection methods and personnel

Project 3 : Rivonia Trial Audio

- The trial that put Nelson Mandela into prison for 27 years
- Recordings made on Dictabelts
- There are no readers or migration technologies available in South Africa
- Requirement: facilitate the process of digitisation
- **Provenance Challenge :** authenticity of originals (often unlabelled) and the methods for digitisation

Common Challenges

- Information to be recorded concerning history of digital objects + traceability back to non-digital artefacts + intangible heritage
- Growth of digital storage – means that multiple copies of everything are stored in many places – which is the "original authentic" version and how can we PROVE THIS
- Difference institutions have different approaches to repository management

Questions for Digital Provenance

1. RECORDING: How to document provenance for digital objects? Digital objects are subject to many processes – how to record these in the history of the object?
2. PROTECTION: How to ensure that the digital object and its provenance information are protected against misuse, change, illegal copying?

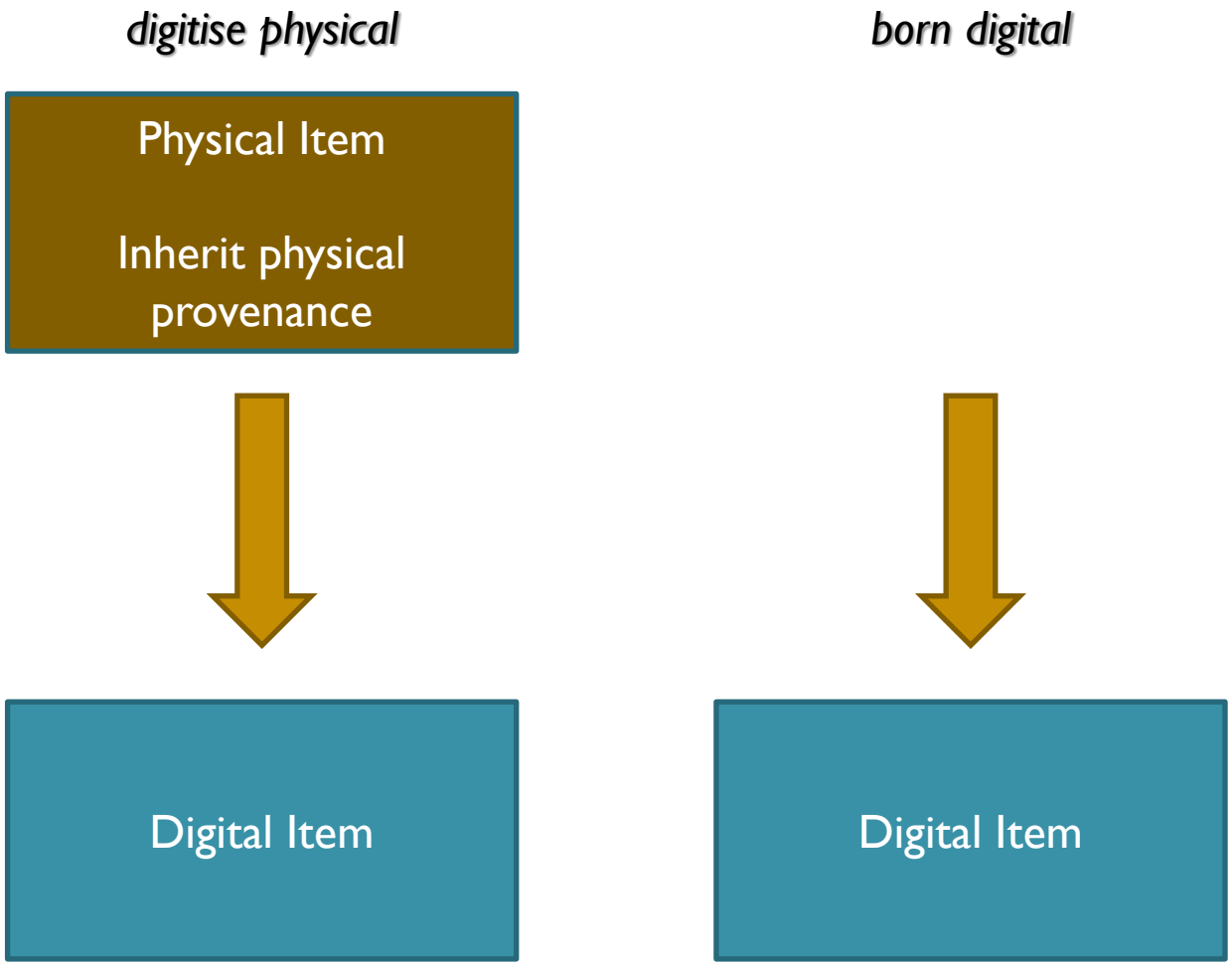
Question 1 : RECORDING

- To document the provenance of a digital objects – means to record everything from the time it was created until it reached its current status.
- May have many versions of the same objects – with different processes (e.g. different resolutions / sizes of an image) which share a common history.

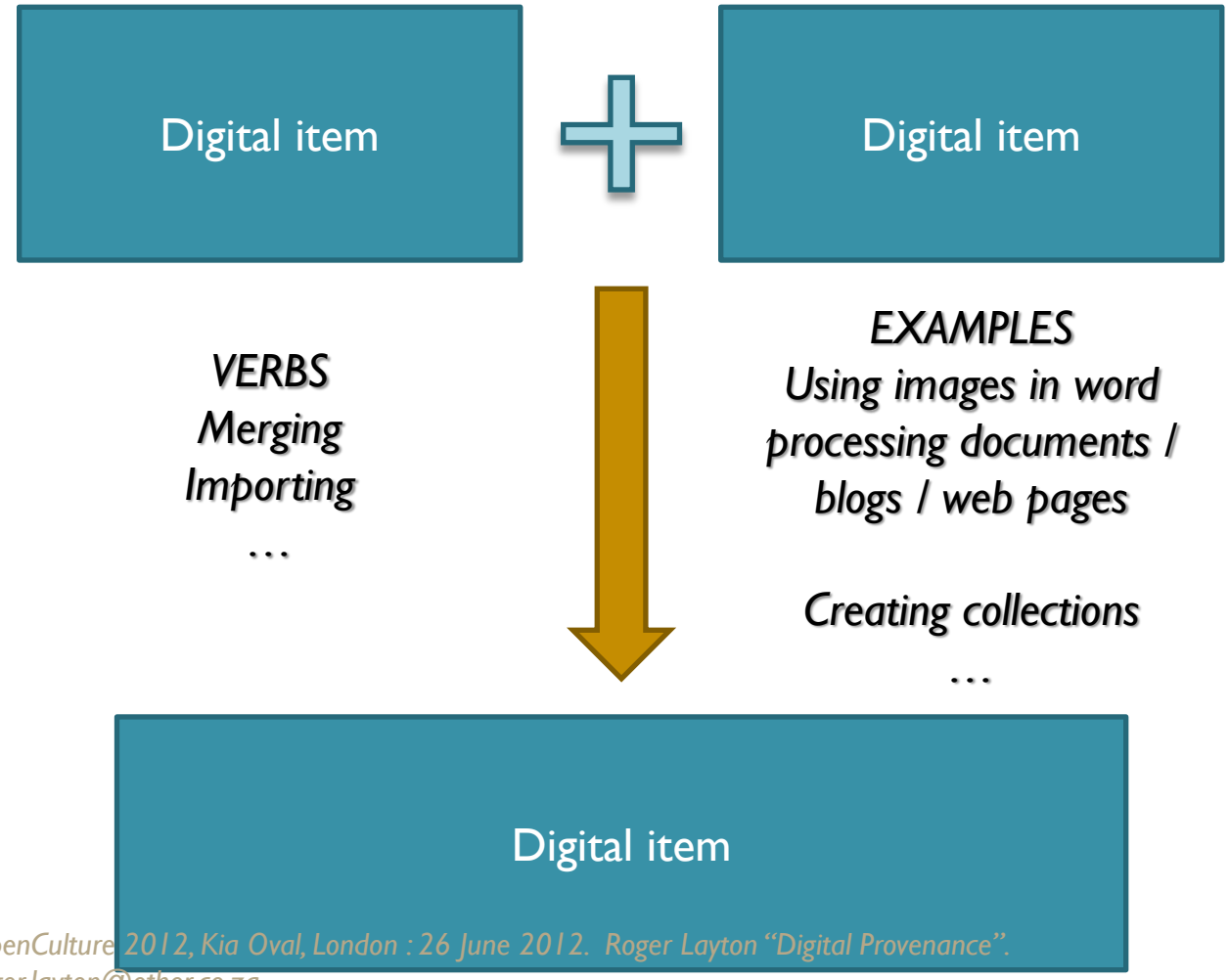
Chemical Engineering Analogy

- What do chemical engineers do?
  - MIX things together
  - FILTER things apart
  - TRANSFORM things into other things
  - MOVE things from one place to another
- What if these "things" are digital objects
- How much do these processes change?

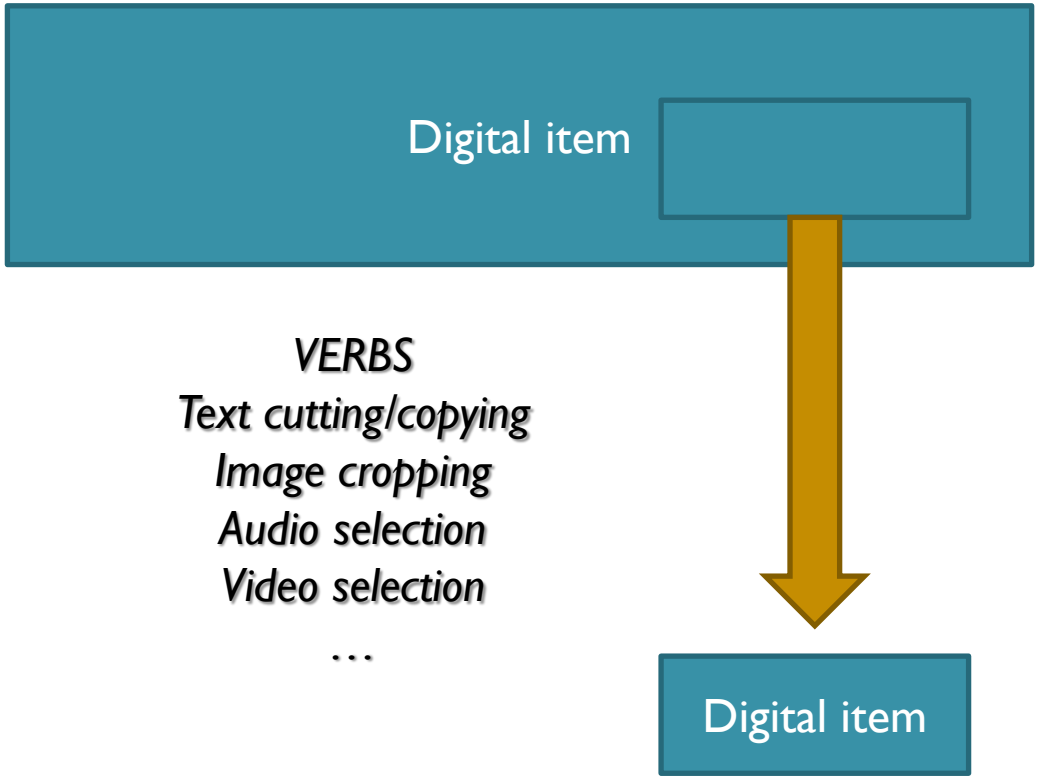
Process 1 : Creating



Process 2 : Packaging (Mixing)

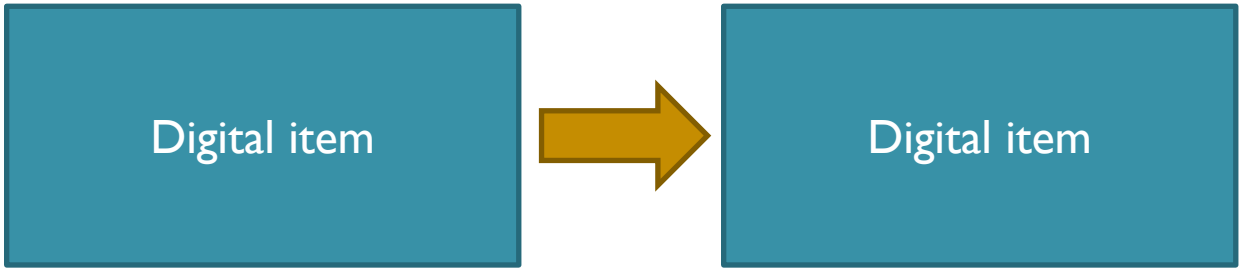


Process 3 : Extracting (Filtering)



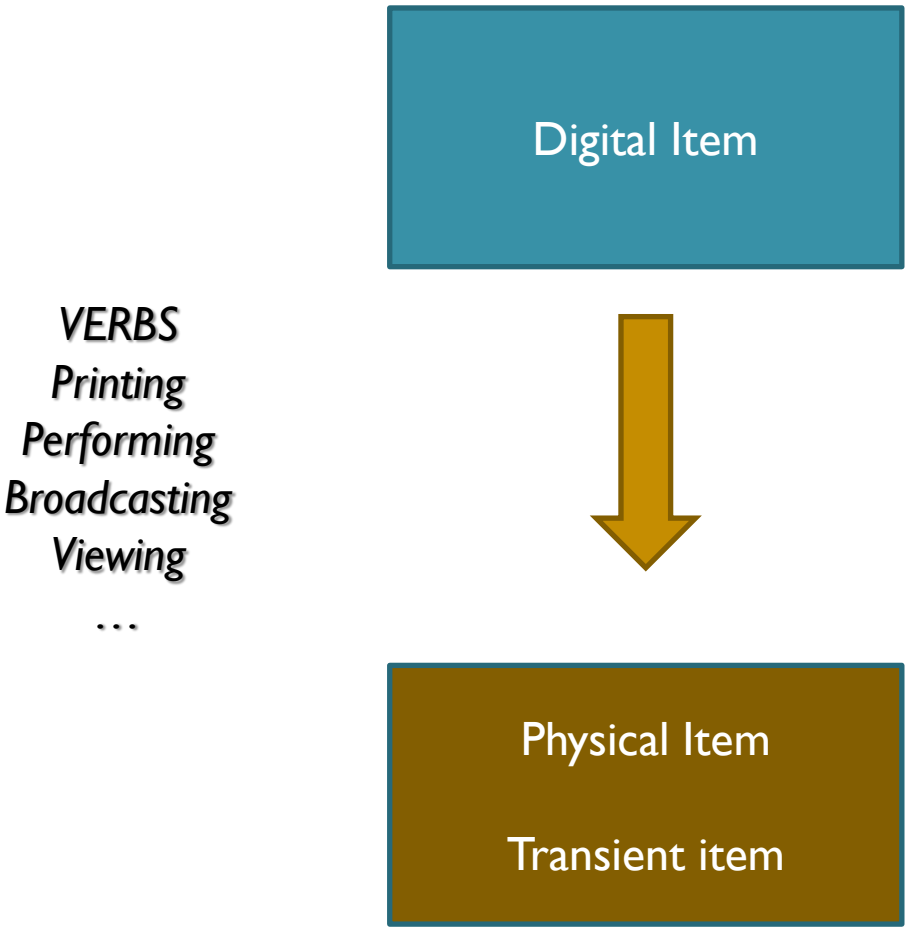
ETHER

Process 4 : Transforming



VERBS
Translation
Resizing
Reformatting
Image processing
Audio/Video processing
Editing
…

Process 5 : Outputting



roger.layton@ether.co.za

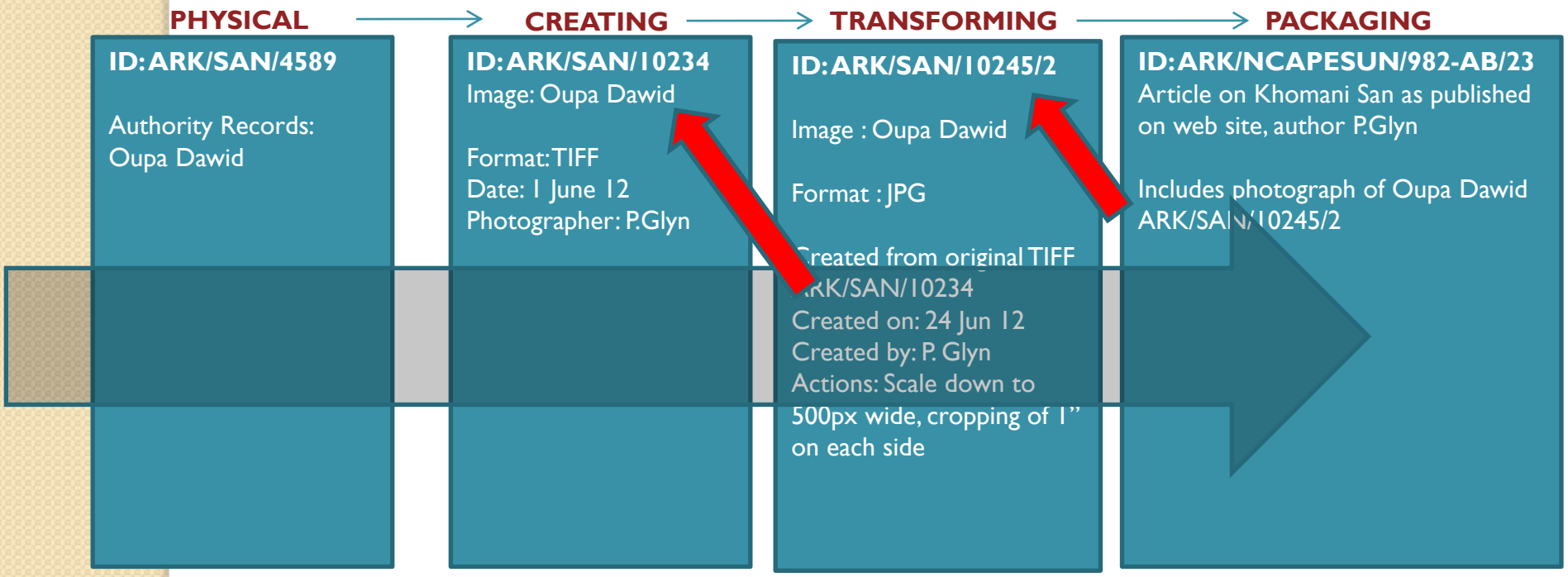
ETHER

Identification of Digital Objects

- Requirements
  - EVERY digital object cited within a digital provenance record must be identified and located (authenticity of reference, metadata and the actual digital object)
  - very long term (eternal) persistence of identifiers : such as the ARK Identifier Scheme (archival resource key)
  - very long term (eternal) persistence of the reference institutions that provide ARK information and house the repositories

How is this recorded?

- In linked graph structures – similar to recommendations of Open Provenance Model
- Similar to hyperlinking in the Web but linking to other provenanced objects
- Every objects within the entire provenance record must be available and stored somewhere in the world's repositories



QUESTION 2 : PROTECTION

- How to protect digital objects against misuse, change, illegal copying?
- How to package all information into a structural unit which cannot be broken up without breaking the objects it contains?

"Digital Masters"

- A concept introduced into the National Policy on Digitisation for South Africa
- Digital Master = a self-contained and protected package
  - Contains digital objects + metadata + any other annotations as required
  - Needs authorised access for different operations (open, view, extract)
  - Signed digitally by the creator
  - Contains total provenance information
  - Contains rights information
  - Contains preservation information

roger.layton@ether.co.za

ETHER

OAIS Information Package

- The "Digital Master" is a form of an Information Package as outlined in the OAIS standard
- Whereas OAIS is a reference model – the Digital Master is an implementation of this model for practical and widespread usage
- EXAMPLE: Passing government archival records to the National Archives – packaging them into Digital Master first to establish contents (no more, no less), authenticity (the right objects) and provenance (how they were created)

How our work continues…

- Building the Digital Provenance model into our ETHER Base product
- Developing standards and practices for the handling of digital content
  - Analogous to SPECTRUM which is directed at physical objects
- Application to heritage / memory institutions in Southern Africa and Africa as a whole

Sustainability of heritage institutions

- PROBLEM:
  - How can small institutions, particular in Africa, survive with less funds available?
  - Current research : how can the digital heritage provide additional income streams
  - We would welcome your input on this! (we are available Thursday/Friday to meet while we are in London)

Summary

- Digital Heritage will be the predominant form of heritage in 20-50 years
- The emerging problem of the storage explosion – total world storage increasing by a factor of 1000 every 10 years
- There is an emerging problem of loss of provenance – a universe of orphans
- We need to solve this project in this generation – leave a good legacy for the next generation