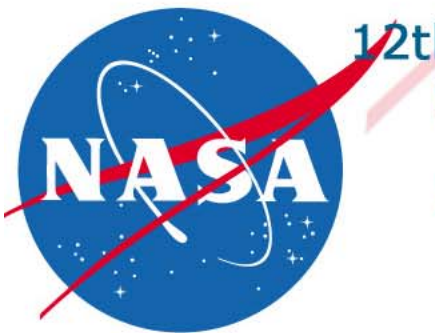
Long-Term Stewardship of Globally-Distributed Representation Information

David Holdsworth

Leeds University

ecldh@leeds.ac.uk

NASA/IEEE MSST 2004
12th NASA Goddard/21st IEEE Conference on
Mass Storage Systems & Technologies
The Inn and Conference Center
University of Maryland University College
Adelphi MD USA
April 13-16, 2004



Long-Term Stewardship of Globally-Distributed Representation Information

David Holdsworth

and

Paul Wheatley

Time-Scale

- **Long-Term**
- 10 years ?
- 100 years ??
- 1000 years ???
- **Beyond current technology**

Summary

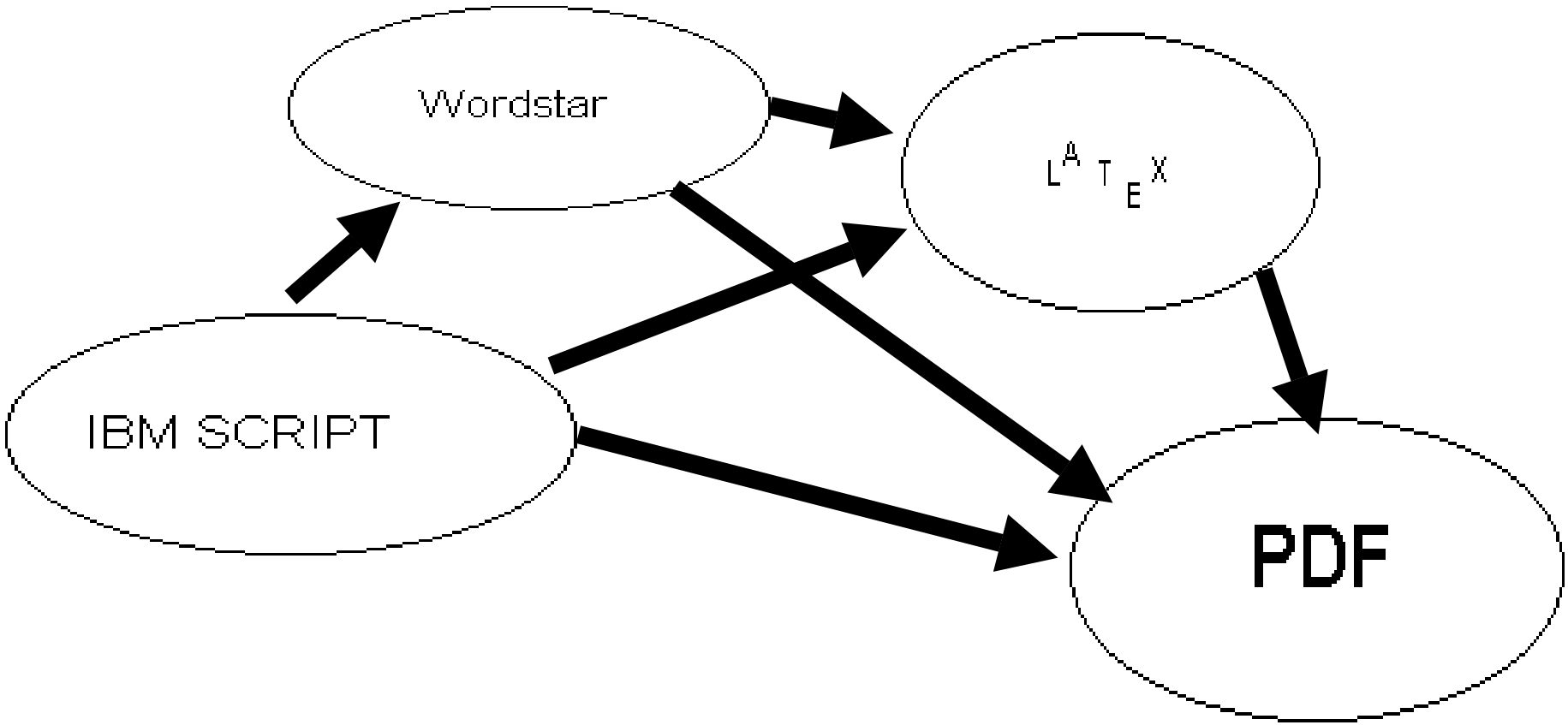
- Construct a byte-stream from which the original's significant properties can be reproduced
- The master is this byte-stream **for ever**
- Maintain indefinitely the ability to access its intellectual content
- Share this knowledge globally
- Change is inevitable **and unpredictable**
- Adaptability and Upwards Compatibility

OAIS

- ISO standard
- Generic Model
- Introduces Representation Nets
- Revision is inevitable
- Avoid overly prescriptive implementations

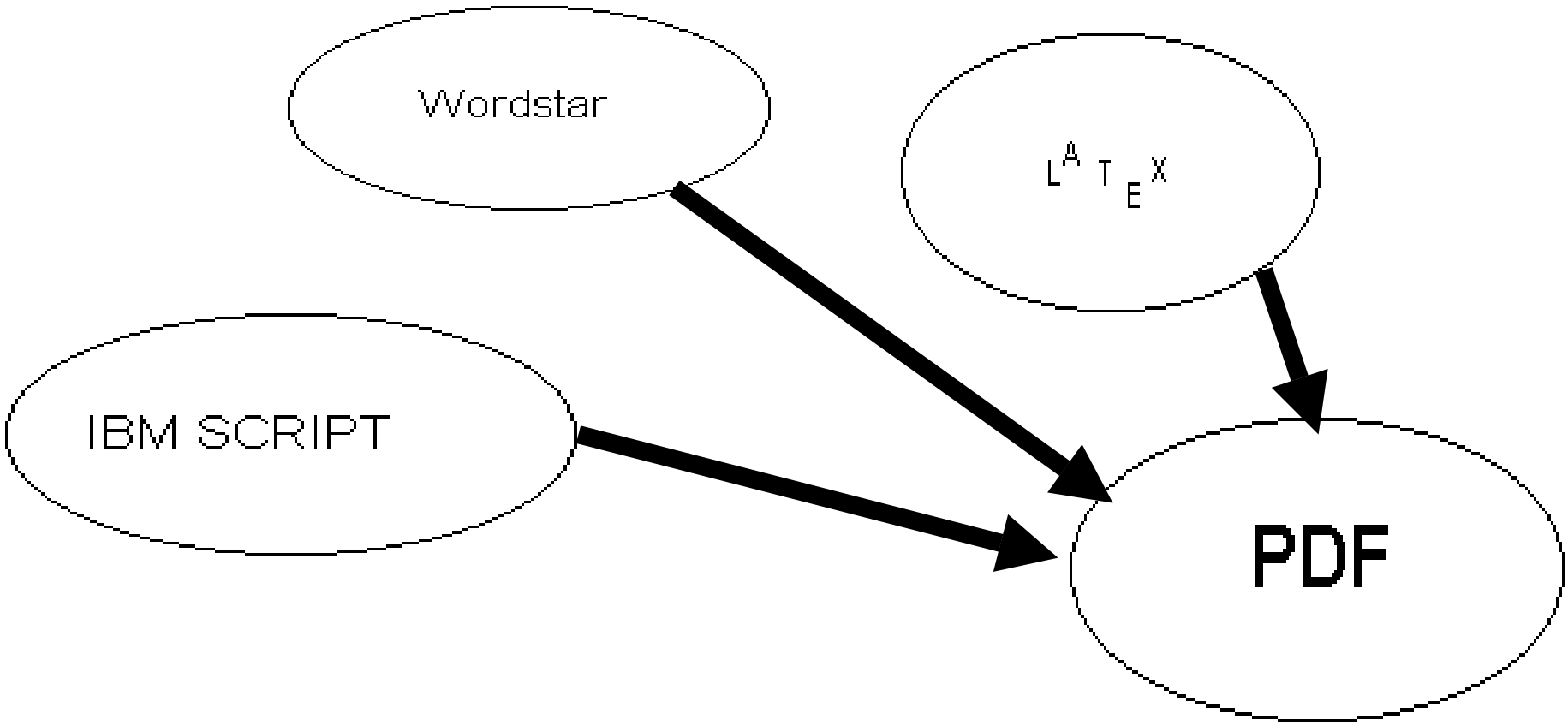
Bridge between Past and Present

- In time future becomes the present



Bridge between Past and Present

- Always we need a bridge to the present



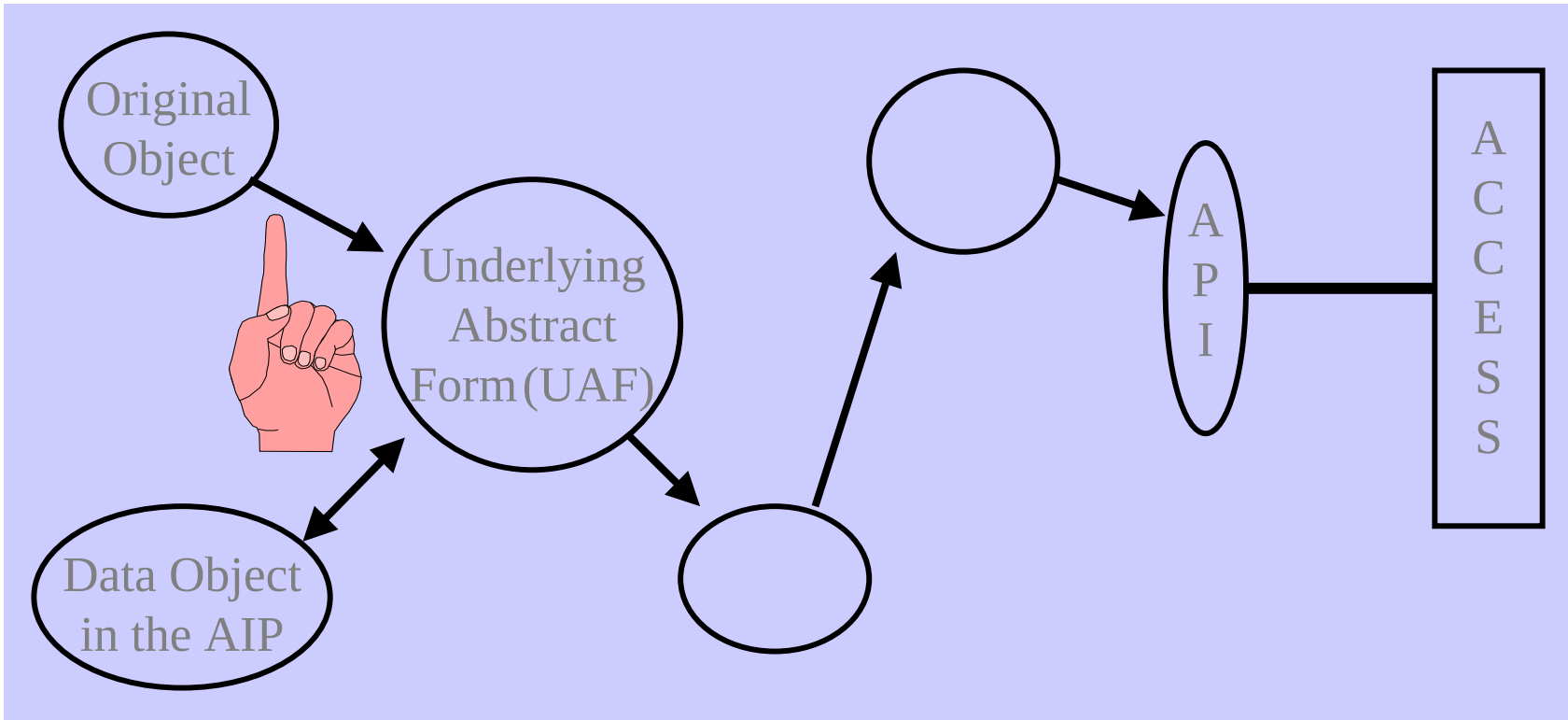
The Future is another country of which we know little

- … but we will get there eventually ….
- ……. or our children will

Abstraction

- What will survive?
- Abstract concept of information
- Abstract concept of **digital** information
- Abstraction enables separation of media and representation issues
- Byte-streams can be stored for ever
  - as we said in MSST2000 …….

Ingest 1

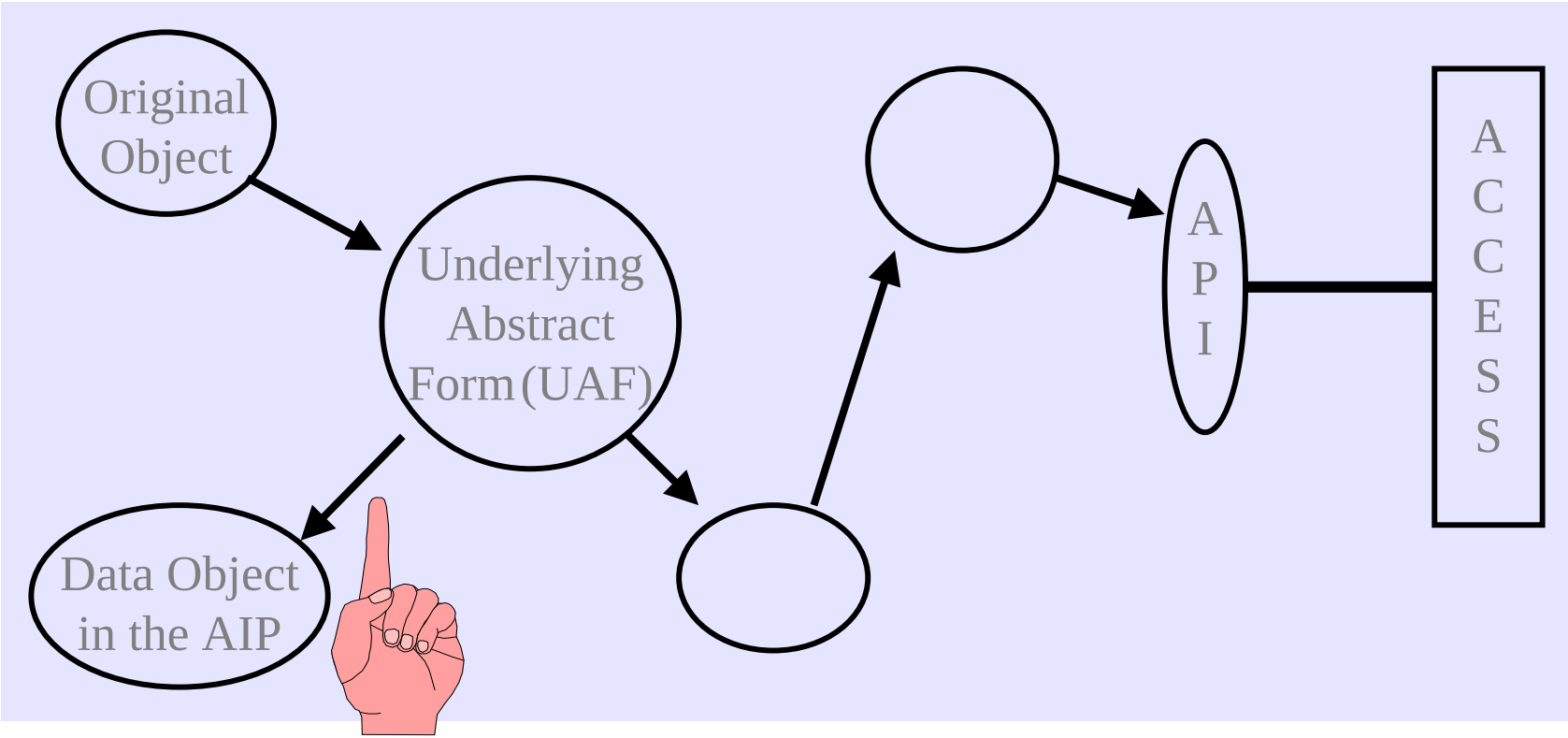


Detach the data from the medium

Underlying Abstract Form - UAF

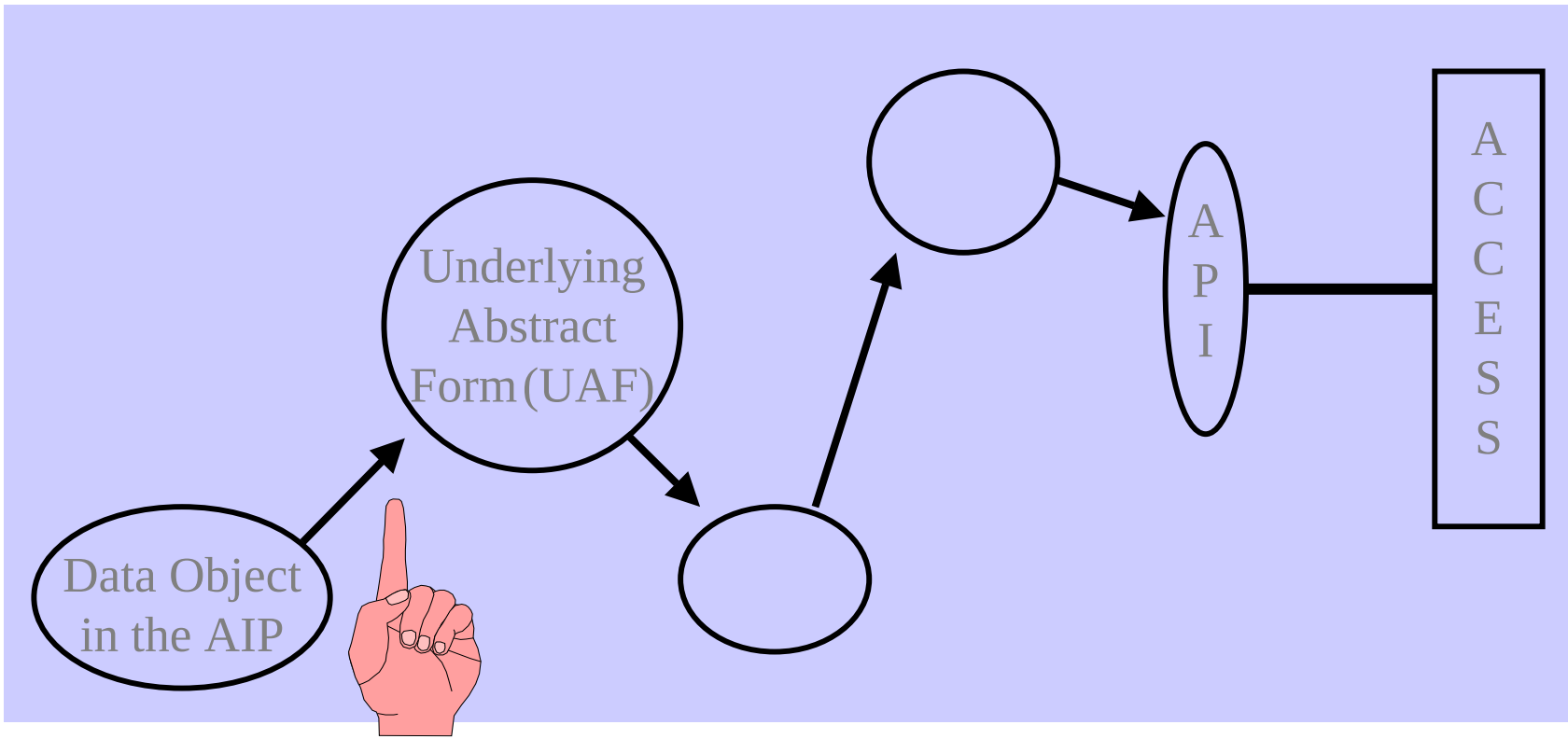
- The ***UAF*** is chosen to preserve the significant properties of the data set
- Identification of ***significant properties*** is vital
- At ingest the data set is mapped to a byte stream

Ingest 2



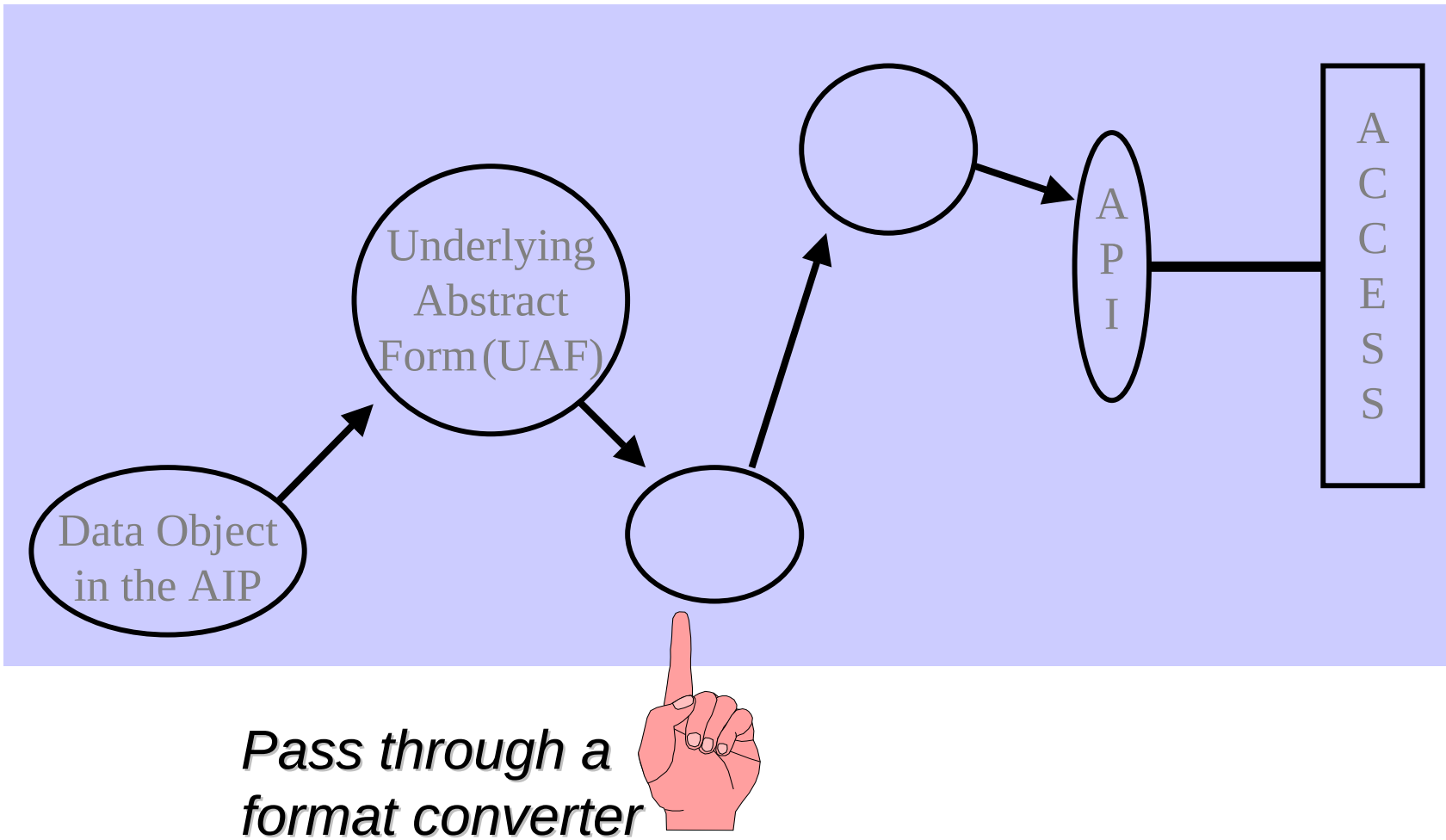
Convert to a byte stream for long-term storage in an Archive Information Package (AIP)

Access 1

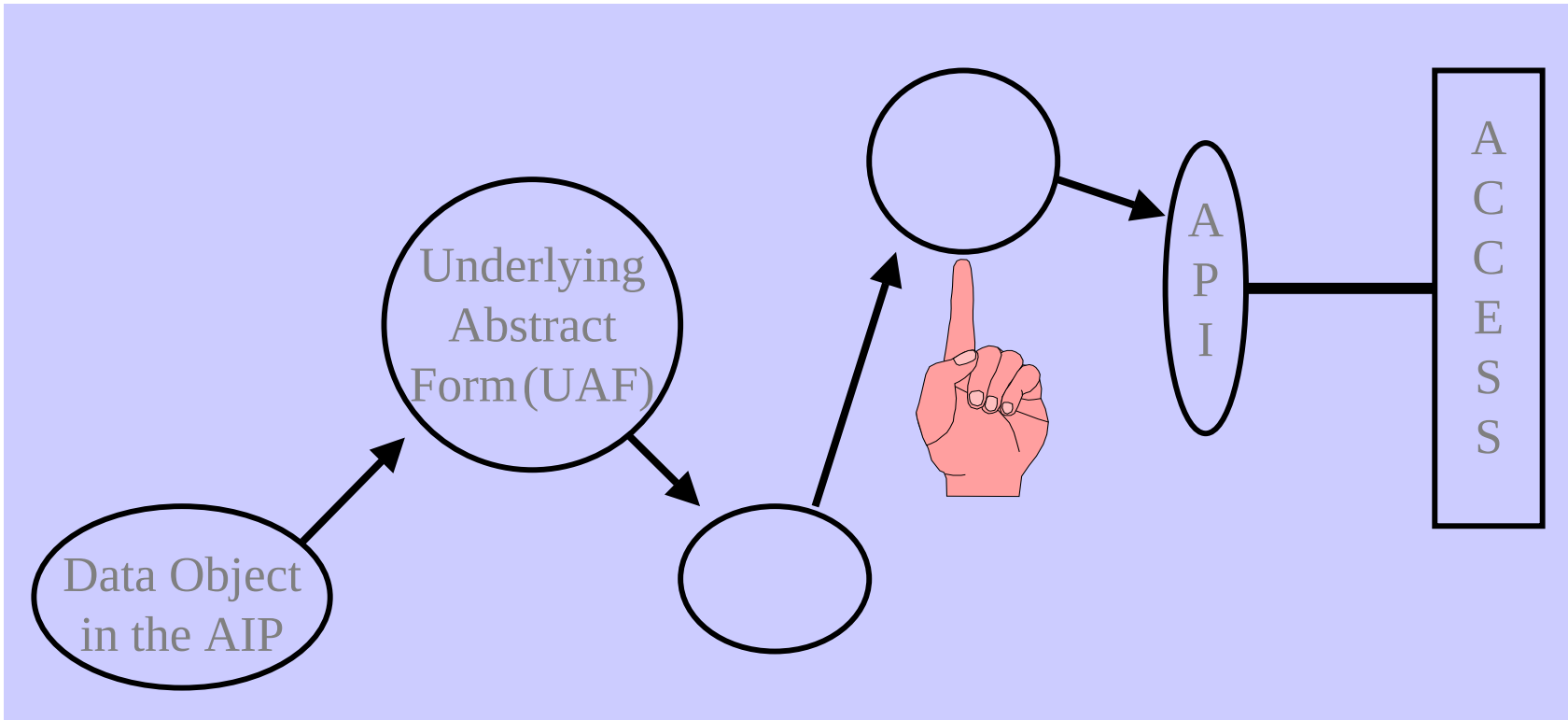


Rebuild the UAF

Access 2

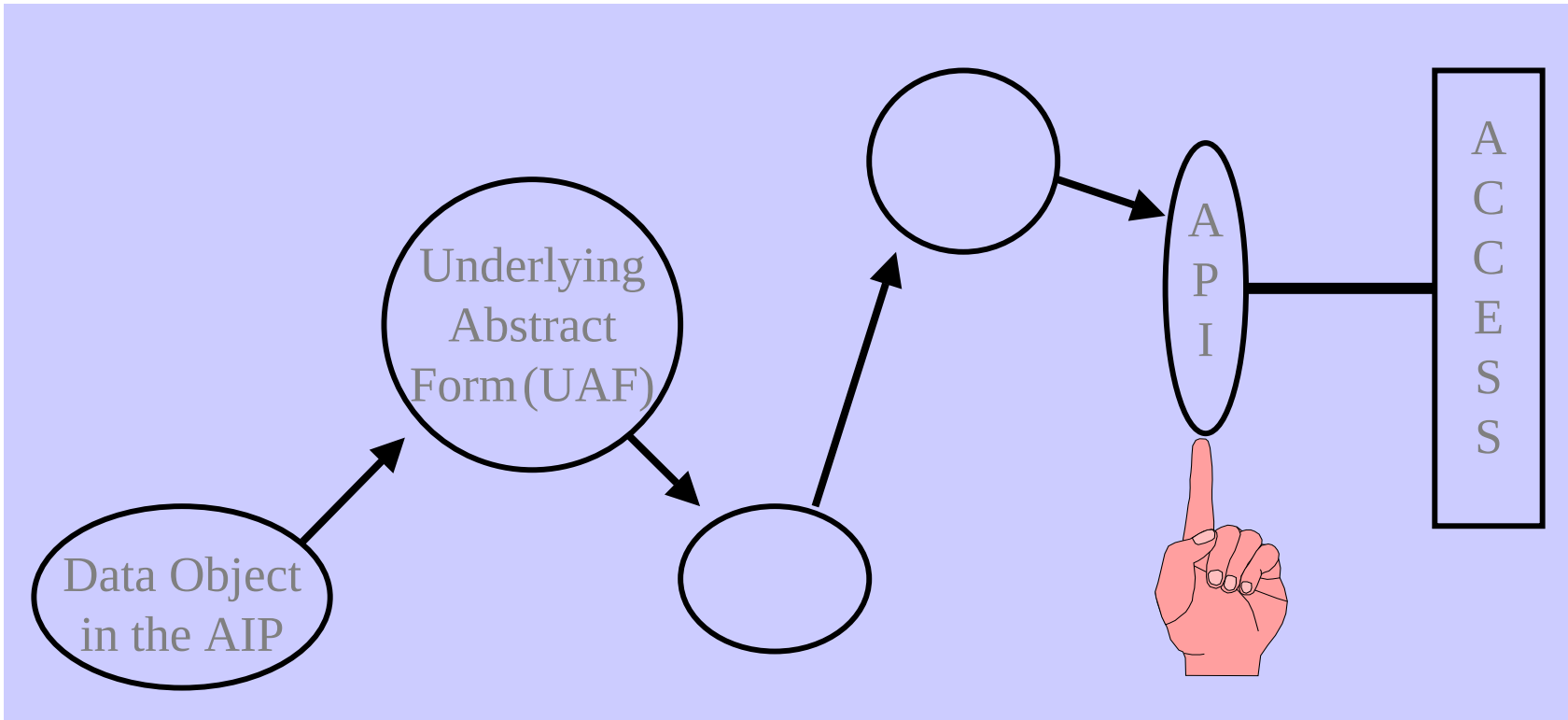

Underlying
Abstract
Form (UAF)
Data Object
in the AIP
API
ACCESS
Pass through a
format converter

Access 3



……. possibly many times ……

Access 4



… to provide an API suitable for access to the intellectual content

Keep the Original

- Always !!
- Bridge gets longer
- Rep net needs monitoring
- Rep net needs maintenance

BBC Domesday

- 1066 William does a bit of conquering
- 1086 Domesday Book
  — Inventory of England
- 1986 Domesday Book still exists
  — Held by British Library
- 1986 BBC makes “modern” version
  — to mark the 900th Anniversary
- 2086 1000th Anniversary

Faith in the Medium

?

Faith in the Technology

Abstraction (UAF) for the Domesday Disc

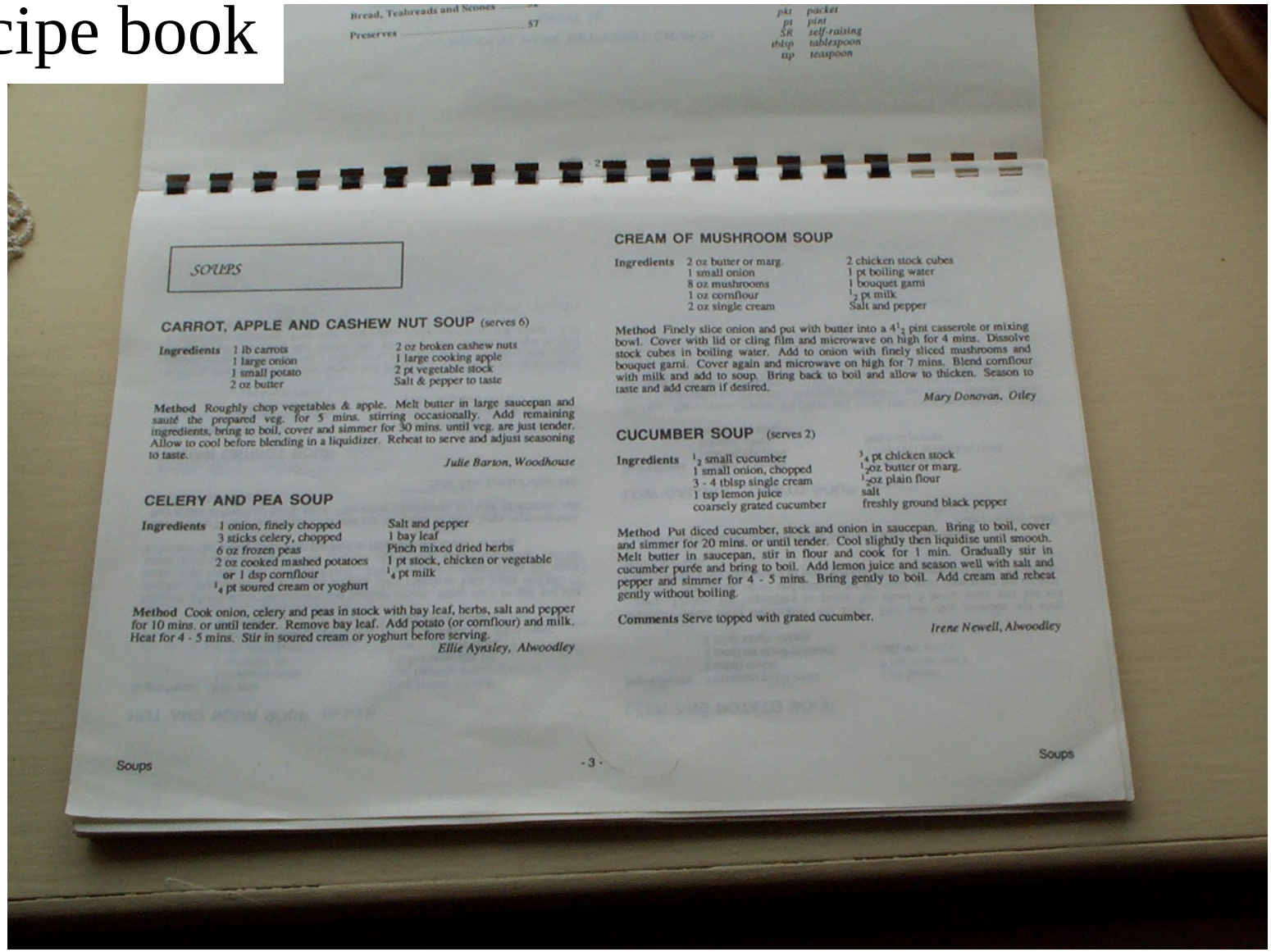
- Each track on the disk is imaged as a single file
- Firmware imaged in files
- Emulator in C allows access to intellectual content on today's platforms

Rep Net Evolution

- Current data needs one node
  - addressed by file extension in MS world view
  - click on a .PDF file and Acrobat reader is invoked automatically
- Versioning
  - often upwards compatible
- Obsolescence of format or current platform
  - time to update the representation net

IBM SCRIPT example

Recipe book



Bread, Teabreads and Scones 2
Preserves 57

pkt packet
pt pint
SR self-raising
tblsp tablespoon
tsp teaspoon

SOUPS

CARROT, APPLE AND CASHEW NUT SOUP (serves 6)

Ingredients 1 lb carrots
1 large onion
1 small potato
2 oz butter
2 oz broken cashew nuts
1 large cooking apple
2 pt vegetable stock
Salt & pepper to taste

Method Roughly chop vegetables & apple. Melt butter in large saucepan and sauté the prepared veg. for 5 mins. stirring occasionally. Add remaining ingredients, bring to boil, cover and simmer for 30 mins. until veg. are just tender. Allow to cool before blending in a liquidizer. Reheat to serve and adjust seasoning to taste.

Julie Barton, Woodhouse

CELERY AND PEA SOUP

Ingredients 1 onion, finely chopped
3 sticks celery, chopped
6 oz frozen peas
2 oz cooked mashed potatoes
or 1 dsp cornflour
¼ pt soured cream or yoghurt
Salt and pepper
1 bay leaf
Pinch mixed dried herbs
1 pt stock, chicken or vegetable
¼ pt milk

Method Cook onion, celery and peas in stock with bay leaf, herbs, salt and pepper for 10 mins. or until tender. Remove bay leaf. Add potato (or cornflour) and milk. Heat for 4 - 5 mins. Stir in soured cream or yoghurt before serving.

Ellie Aynsley, Alwoodley

CREAM OF MUSHROOM SOUP

Ingredients 2 oz butter or marg.
1 small onion
8 oz mushrooms
1 oz cornflour
2 oz single cream
2 chicken stock cubes
1 pt boiling water
1 bouquet garni
½ pt milk
Salt and pepper

Method Finely slice onion and put with butter into a 4½ pint casserole or mixing bowl. Cover with lid or cling film and microwave on high for 4 mins. Dissolve stock cubes in boiling water. Add to onion with finely sliced mushrooms and bouquet garni. Cover again and microwave on high for 7 mins. Blend cornflour with milk and add to soup. Bring back to boil and allow to thicken. Season to taste and add cream if desired.

Mary Donovan, Otley

CUCUMBER SOUP (serves 2)

Ingredients ½ small cucumber
1 small onion, chopped
3 - 4 tblsp single cream
1 tsp lemon juice
coarsely grated cucumber
¾ pt chicken stock
½oz butter or marg.
½oz plain flour
salt
freshly ground black pepper

Method Put diced cucumber, stock and onion in saucepan. Bring to boil, cover and simmer for 20 mins. or until tender. Cool slightly then liquidise until smooth. Melt butter in saucepan, stir in flour and cook for 1 min. Gradually stir in cucumber purée and bring to boil. Add lemon juice and season well with salt and pepper and simmer for 4 - 5 mins. Bring gently to boil. Add cream and reheat gently without boiling.

Comments Serve topped with grated cucumber.

Irene Newell, Alwoodley

IBM SCRIPT example

The “intellectual” content

SOUPS

CARROT, APPLE AND CASHEW NUT SOUP (serves 6)

Ingredients
1 lb carrots
1 large onion
1 small potato
2 oz butter
2 oz broken cashew nuts
1 large cooking apple
2 pt vegetable stock
Salt & pepper to taste

Method Roughly chop vegetables & apple. Melt butter in large saucepan and sauté the prepared veg. for 5 mins. stirring occasionally. Add remaining ingredients, bring to boil, cover and simmer for 30 mins. until veg. are just tender. Allow to cool before blending in a liquidizer. Reheat to serve and adjust seasoning to taste.

Julie Barton, Woodhouse

CELERY AND PEA SOUP

Ingredients
1 onion, finely chopped
3 sticks celery, chopped
6 oz frozen peas
2 oz cooked mashed potatoes
or 1 dsp cornflour
¼ pt soured cream or yoghurt
Salt and pepper
1 bay leaf
Pinch mixed dried herbs
1 pt stock, chicken or vegetable
¼ pt milk

Method Cook onion, celery and peas in stock with bay leaf, herbs, salt and pepper for 10 mins. or until tender. Remove bay leaf. Add potato (or cornflour) and milk. Heat for 4 - 5 mins. Stir in soured cream or yoghurt before serving.

Ellie Aynsley, Atwoodley

CREAM OF MUSHROOM SOUP

Ingredients
2 oz butter or marg.
1 small onion
8 oz mushrooms
1 oz cornflour
2 oz single cream
2 chicken stock cubes
1 pt boiling water
1 bouquet garni
½ pt milk
Salt and pepper

Method Finely slice onion and put with butter into a 4½ pint casserole or mixing bowl. Cover with lid or cling film and microwave on high for 4 mins. Dissolve stock cubes in boiling water. Add to onion with finely sliced mushrooms and bouquet garni. Cover again and microwave on high for 7 mins. Blend cornflour with milk and add to soup. Bring back to boil and allow to thicken. Season to taste and add cream if desired.

Mary Donovan, Otley

CUCUMBER SOUP (serves 2)

Ingredients
½ small cucumber
1 small onion, chopped
3 - 4 tblsp single cream
1 tsp lemon juice
coarsely grated cucumber
¾ pt chicken stock
½oz butter or marg.
½oz plain flour
salt
freshly ground black pepper

Method Put diced cucumber, stock and onion in saucepan. Bring to boil, cov and simmer for 20 mins. or until tender. Cool slightly then liquidise until smoo Melt butter in saucepan, stir in flour and cook for 1 min. Gradually stir cucumber purée and bring to boil. Add lemon juice and season well with salt pepper and simmer for 4 - 5 mins. Bring gently to boil. Add cream and reh gently without boiling.

Comments Serve topped with grated cucumber.

Irene Newell, Alwoo

IBM SCRIPT example

- Digital Original

use **SCRIPT** command on VM/CMS

```
.IM APHEAD
.BF title
.BT /Soups/- % -/Soups/
.BX ON 1 30
 .BI
    SOUPS
.SU;.PU 1 +&syspage+ SOUPS

.BX OFF
.PF
 .CC 12;.BI CARROT,  APPLE  AND  CASHEW  NUT  SOUP\  (serves 6)
.SU;.PU 1 -&syspage- CARROT,  APPLE  AND  CASHEW  NUT  SOUP\  (serves 6)

.BD Ingredients\|1 lb carrots|2 oz broken cashew nuts
 |1 large onion|1 large cooking apple
 |1 small potato|2 pt vegetable stock
 |2 oz butter|Salt & pepper to taste

.BD Method\  Roughly chop vegetables & apple.  Melt butter in large
saucepan and saute the prepared veg. for 5 mins. stirring
occasionally.  Add remaining ingredients, bring to boil, cover and
simmer for 30 mins. until veg. are just tender.  Allow to cool before
blending in a liquidizer.  Reheat to serve and adjust seasoning to
taste.
.RI;.US Julie Barton, Woodhouse
```

Hercules to the rescue

(emulates IBM/370 on PC)

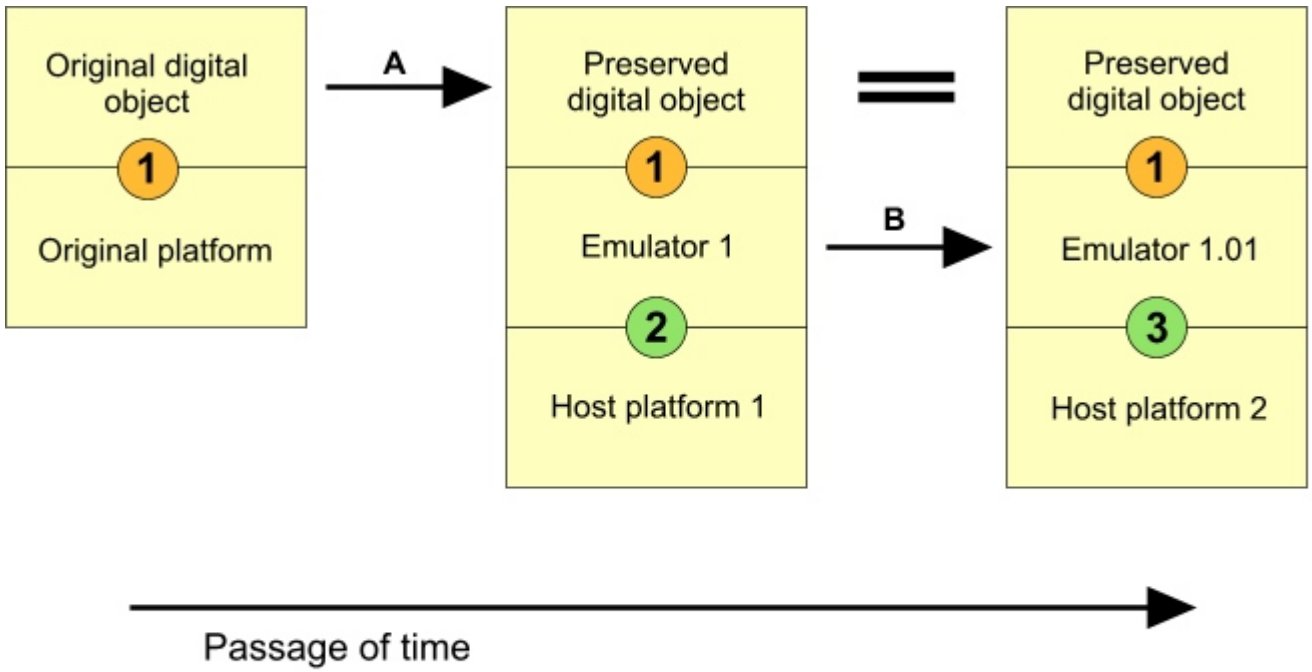
Indirection

- Update of rep net
  - takes account of passage of time
- Rep net referenced from archived data (AIP)
- Lots of objects reference same rep net
- Persistent naming
  - need a system of reference Ids
- Modify rep info to include Hercules

Software and rep nets

- Rep net must enable access to intellectual content
  - This needs software
- Embedded or Links?
- Keep the software as legitimate digital data
- Rep net must record platforms
- Emulation provides recreation of original
- Users also want to use current tools

Software Evolution



Emulator written to long-lived API

Abstract Emulation Interface

Abstract Emulation Platform

File format registry(ies)

- Global File Format Registry
  - Harvard
- PRONOM
  - National Archives in the UK
- Can be referenced from rep nets

Sharing of rep nets

- Enabled if we have global persistent names
- Encouraged if information is reliable
- Economies of scale

Referential Integrity
Avoiding Broken Links

- Distributed
  - minimising effect of major disasters
- Replicated
  - recovering from major disasters
- Resolution
  - translating permanent name to current location
- Naming Authorities
  - ensuring uniqueness

LOCKSS

- Lots
- Of
- Copies
- Keep
- Stuff
- Safe

Stanford University
and
Sun Microsystems

Traditional
approach
of
libraries

Ongoing format migration
why not?

- Keep the data in current formats
  - involves repeated format conversion
- Error build-up
  - caused by mismatch between formats
  - undetected errors in conversion process
- Only sensible when doing media change
  - couples format and media issues

Global Picture

- Not everyone agrees
  - ***"Be reasonable. Do it my way"***
- Not everyone has exactly the same needs
- Co-operation between equals
- Need to accommodate multiple approaches
- Good discipline for future-proofing

Authorities

- DCC - UK's new Digital Curation Centre
- OAIS – ISO
- CCSDS
- NASA
- Co-operation and consensus
- not competition

Curating the Future

Peter Burnhill

Director (Phase One), DCC

(Director, EDINA National Data Centre)

Funded by: JISC e-Science

UK Digital Curation Centre

- Call made in JISC Circular 6/03, joint funding
  - by the JISC and the e-Science Core Programme
    - Funding for outreach, services & development
    - Funding for research programme
    - Challenge of ensuring fruitful interaction
- Twin drivers
  - e-Science - ‘data deluge’ - continuing access
  - Digital Preservation
- Ambitious & demanding remit
  - across disciplines (research councils), HE, FE and beyond
  - across data types
- Task entrusted to Consortium of four partners

D | C | C

DCC Consortium Partners

- Four Consortium partner institutions:
  - University of Edinburgh - lead partner
  - University of Glasgow (HATII)
  - University of Bath (UKOLN)
  - CCLRC (Rutherford and Daresbury Laboratories)

- Prior links via National eScience Centre (NeSC)
  - jointly managed by Universities of Edinburgh & Glasgow

Overall Aim

‘continuing quality improvement in data curation & digital preservation’

Initial focus:

data as evidential base for scholarly conclusions

- role of data archiving & preservation as keys to reproducibility and reuse

Wider context & remit:

worlds of scholarly communication & eLearning

DCC

Objectives

- vibrant research programme
  - addressing the wider issues of digital curation
- Collaborative Associates Network of Data Organisations
  - strong links across existing community of practice
  - engagement with curators (individuals & organisations)
- services
  - to evaluate tools, methods, standards and policies
  - a repository of tools and technical information
- ‘virtuous circle’
  - expertise, experience & requirement feed into the DCC research programme

DCC

Research & Development

- Research
  - Annotation, Data integration and publication
  - Appraisal and long-term preservation
  - Socio-economic & legal context
    - rights, responsibilities and viability
  - Performance and Optimisation
- Development into Services
  - Standards & Testbeds
  - File Formats
  - Registry of Metadata Standards

Further topics:

- Evolution of structure, Ontologies, Emulation

Data Curation & Terminology

- actions needed to maintain and utilise digital data & research results over entire life-cycle
  - for current and future generations of users.

alongside which is Archiving

  - appraisal & retention/disposal
    - logical & physical integrity: authenticity/security

and Digital Preservation

  - long-run technological/legal accessibility & usability

- Data curation in science
  - maintenance of body of trusted data
    - to represent current state of knowledge in area of research.

D C C

Long-Term Stewardship of Globally-Distributed Representation Information

David Holdsworth

Leeds University

ecldh@leeds.ac.uk

NASA/IEEE MSST 2004
12th NASA Goddard/21st IEEE Conference on
Mass Storage Systems & Technologies
The Inn and Conference Center
University of Maryland University College
Adelphi MD USA
April 13-16, 2004