

CASPAR
Cultural, Artistic and Scientific knowledge
for Preservation, Access and Retrieval

Preservation Threats

David Giaretta
CASPAR Project Director

Training on preservation of Cultural and Scientific Data

CASPAR

Information is the important thing

Information:
Any type of knowledge that can be exchanged. In an exchange, it is represented by data.

Long Term is long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community. Long Term may extend indefinitely.

Ensure that the information to be preserved is Independently Understandable to (and usable by) the Designated Community.

- What information?
 - Documents.....
 - Data.....
- Original bits?
- Look and feel?
- Behaviour?
- Performance?
- Explicit/ Implicit/ Tacit


CASPAR

Issues of transferring info to future custodians

- Things change:
 - Software
 - Hardware
 - Environment
 - E.g. Network links to related information
 - People
 - What is "common knowledge"
 - Organisations and systems
- Chain of preservation
 - Only as strong as its weakest link

How can we ensure that the information trapped in the "bits" remains **understandable** despite all these changes?

How can current custodian prepare for or even be aware of these changes?



CASPAR

Time is short...

- Neither an individual nor his/her institution (or preservation project) will last forever
- The chain of preservation is only as strong its weakest link
- Need to be prepared to hand over

How can whole collections be handed over?

How can the information in the archive managers' heads be handed over?

CASPAR

No repository is an island

- An organisation/project cannot do everything
 - Things change
 - Nothing will be around forever
- Must somehow tap into other resources

How can we find these resources?

How can we share the resources?

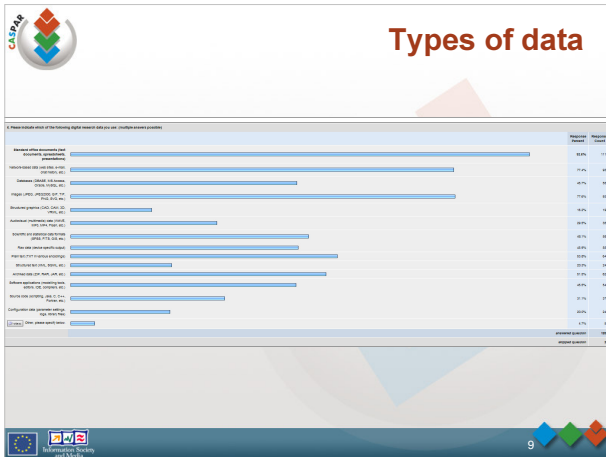
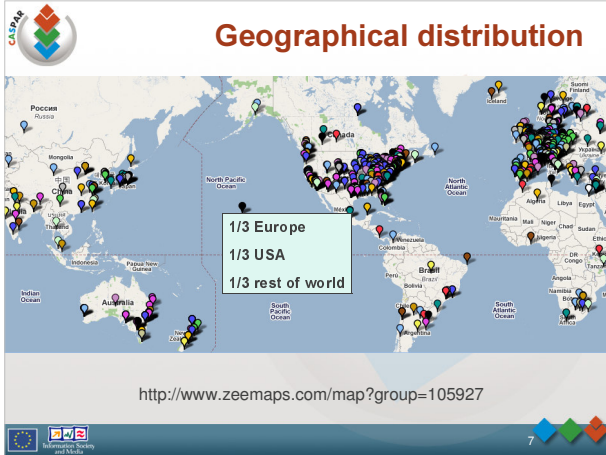
Where do the resources come from?

CASPAR

Survey and preliminary results

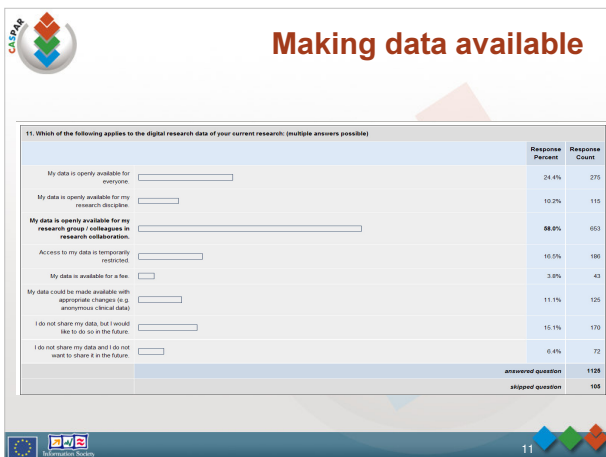
- PARSE.Insight plus Alliance for Permanent Access
- Plus customised surveys with CASPAR, DCC
- Targets
 - Researchers
 - Plus case studies in HEP, (Earth Observation and Social Sciences)
 - Publishers
 - Funders
 - Data managers
- More than 2000 responses so far

- 1) Creation and use of digital research data
- 2) Data Re-use
- 3) Data Preservation
- 4) Publishing Your Work
- 5) Final questions



Types of data

Standard office documents (text documents, spreadsheets, presentations)	92.7%	1,145
Network-based data (web sites, e-mail, chat history, etc.)	77.5%	957
Databases (DBASE, MS Access, Oracle, MySQL, etc.)	45.6%	563
Images (JPEG, JPEG2000, GIF, TIF, PNG, SVG, etc.)	77.9%	962
Structured graphics (CAD, CAM, 3D, VRML, etc.)	16.4%	203
Audiovisual (multimedia) data (WAVE, MP3, MP4, Flash, etc.)	29.6%	365
Scientific and statistical data formats (SPSS, FITS, GIS, etc.)	48.0%	593
Raw data (device specific output)	45.8%	566
Plain text (TXT in various encodings)	53.7%	663
Structured text (XML, SGML, etc.)	20.5%	253
Archived data (ZIP, RAR, JAR, etc.)	51.7%	639
Software applications (modelling tools, editors, IDE, compilers, etc.)	45.3%	559
Source code (scripting, Java, C, C++, Fortran, etc.)	31.1%	384
Configuration data (parameter settings, logs, library files)	20.2%	250
Other, please specify below.	4.8%	59



Is your data available?

My data is openly available for everyone.	24.7%	285
My data is openly available for my research discipline.	10.3%	119
My data is openly available for my research group / colleagues in research collaboration.	58.1%	669
Access to my data is temporarily restricted.	16.8%	194
My data is available for a fee.	3.8%	44
My data could be made available with appropriate changes (e.g. anonymous clinical data)	11.1%	128
I do not share my data, but I would like to do so in the future.	14.8%	171
I do not share my data and I do not want to share it in the future.	6.4%	74



Problems in sharing data

12. Increasingly, awareness is growing that data should be shared as well as publications. Do you experience or foresee any of the following problems in sharing your data? (multiple answers possible)

	Response Percent	Response Count
Fear to lose scientific edge	27.1%	306
Incompatible data types	32.3%	363
Restricted access to data archive	20.4%	230
Legal issues	40.5%	456
Lack of technical infrastructure	27.3%	307
Misuse of data	41.7%	469
Lack of financial resources	20.9%	235
No problems foreseen	10.6%	119
Other (please specify)	9.2%	103
answered question		1126
skipped question		108



Concerns about sharing data...

Fear to lose scientific edge	27.2%	313
Incompatible data types	32.7%	377
Restricted access to data archive	20.6%	237
Legal issues	40.5%	466
Lack of technical infrastructure	27.6%	318
Misuse of data	41.7%	480
Lack of financial resources	26.8%	309
No problems foreseen	16.6%	191
Other (please specify)	9.3%	107



Lack of availability of data

15. Did you ever need digital research data gathered by other researchers that was not available?

	Response Percent	Response Count
Yes	51.6%	571
No	30.0%	332
Don't know	18.4%	203
answered question		1108
skipped question		124



Did you ever need digital research data gathered by other researchers that was not available?

Yes	52.1%	590
No	29.6%	335
Don't know	18.4%	208



Benefits of re-use

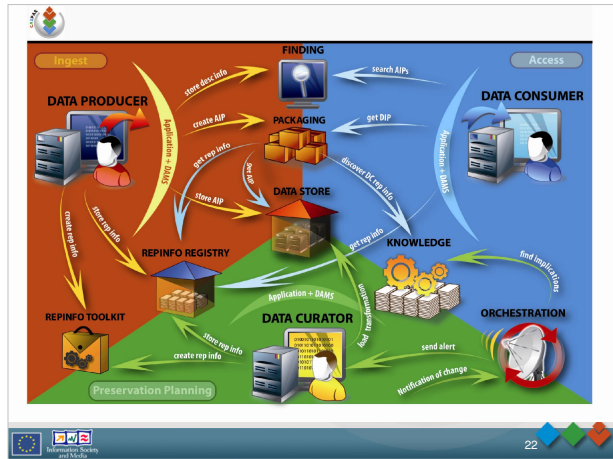
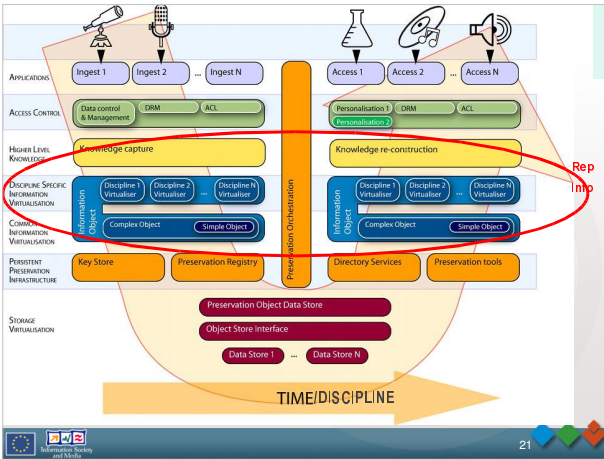
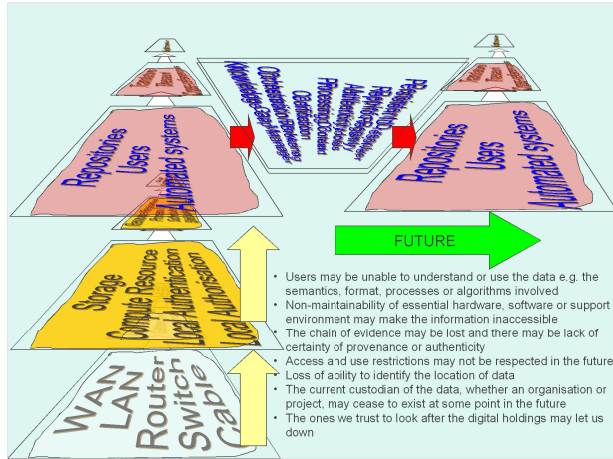
Collaborate in their discipline and across disciplines	55%
Make use of data gathered by other researchers in the discipline	65%
Make use of data from other disciplines	46%
Would like to make use of data from other disciplines	39%
Have needed digital research data gathered by other researchers that was not available	51%
Believe that reuse of data for validation purposes is an important reason for preservation.	80%



Threats to preservation/re-use

	Very important	Important	Slightly important	Not important	Don't know
Users may be unable to understand or use the data e.g. the semantics, format, processes or algorithms involved	33.1% (352)	41.0% (443)	16.6% (177)	6.2% (66)	2.4% (26)
Non-maintainability of essential hardware, software or support environment may make the information inaccessible	40.1% (427)	40.3% (430)	12.9% (138)	5.3% (56)	1.4% (15)
The chain of evidence may be lost and there may be lack of certainty of provenance or authenticity	31.9% (340)	45.4% (484)	16.7% (178)	3.7% (39)	2.3% (25)
Access and use restrictions may not be respected in the future	18.7% (198)	36.9% (391)	24.9% (264)	13.5% (143)	6.0% (64)
Loss of ability to identify the location of data	24.2% (256)	44.5% (470)	22.9% (242)	4.7% (50)	3.6% (38)
The current custodian of the data, whether an organization or project, may cease to exist at some point in the future	35.5% (377)	41.9% (445)	15.8% (168)	4.5% (48)	2.4% (25)

Threat	Requirement for solution
Users may be unable to understand or use the data e.g. the semantics, format, processes or algorithms involved	Ability to create and maintain adequate Representation Information
Non-maintainability of essential hardware, software or support environment may make the information inaccessible	Ability to share information about the availability of hardware and software and their replacements/substitutes
The chain of evidence may be lost and there may be lack of certainty of provenance or authenticity	Ability to bring together evidence from diverse sources about the Authenticity of a digital object
Access and use restrictions may make it difficult to reuse data, or alternatively may not be respected in future	Ability to deal with Digital Rights correctly in a changing and evolving environment
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	Certification process so that one can have confidence about whom to trust to preserve data holdings over the long term



Conclusions

- We must work together on many issues if we wish to preserve the digitally encoded information on which we are increasingly dependent
- CASPAR is developing solutions which are being tested using a variety of real data from science, culture and contemporary performing arts.

END