Digital Curation of Research Data

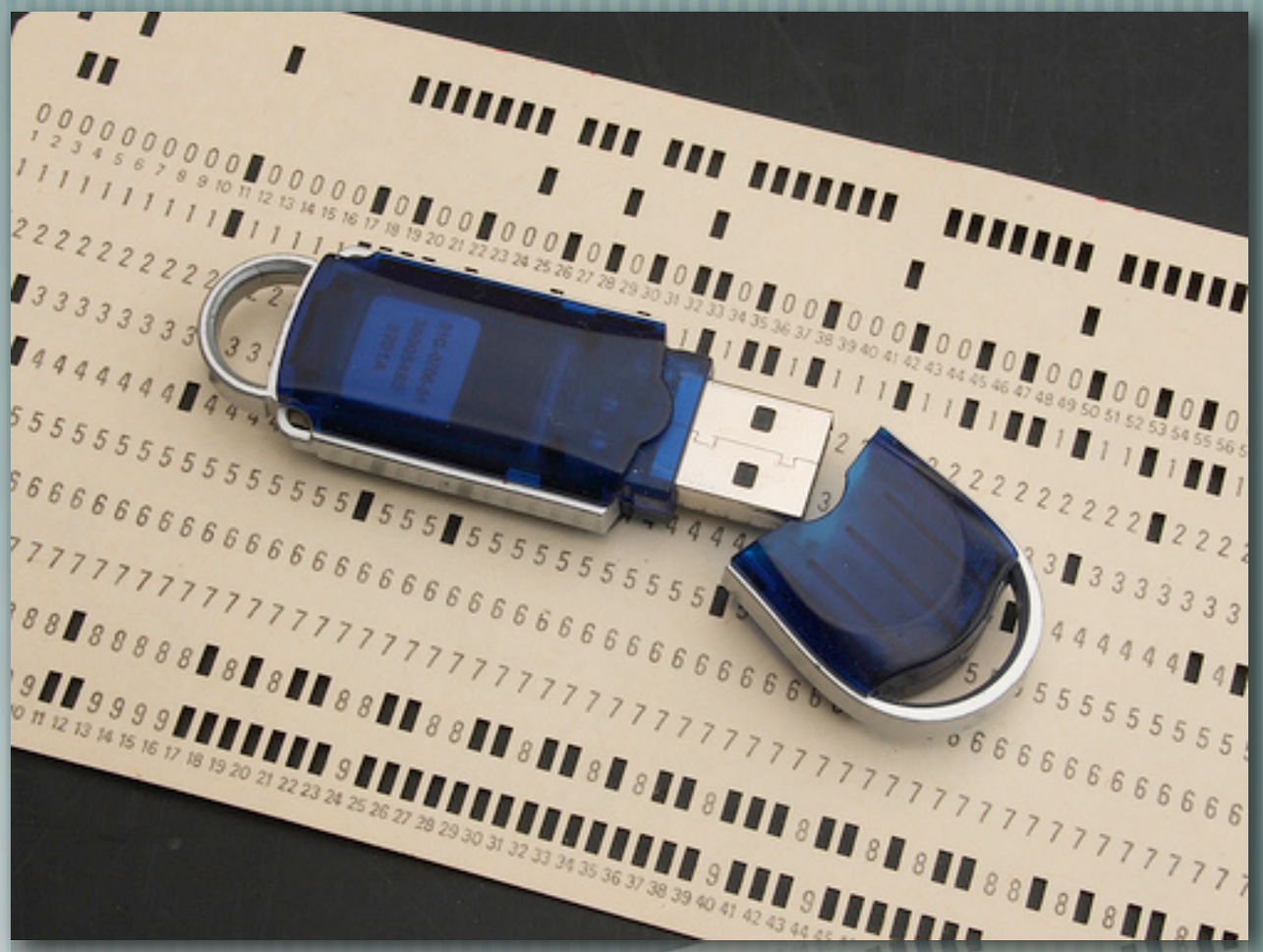
Understanding the lifecycle of dataset preservation

Presented by:

Isaiah Beard

Digital Data Curator

Rutgers University Libraries



March 30 2012 • Rutgers University Libraries

Wednesday, July 11, 12

- This presentation is about crystalizing concepts we're all pretty sure we know, but haven't really addressed formally.
- Research data requires stepping out of our comfort zones: known formats and processes aren't always going to help us in every case, and we'll need to be ready to adapt.
- This presentation will help get us into that mindset.

Topics Covered Today

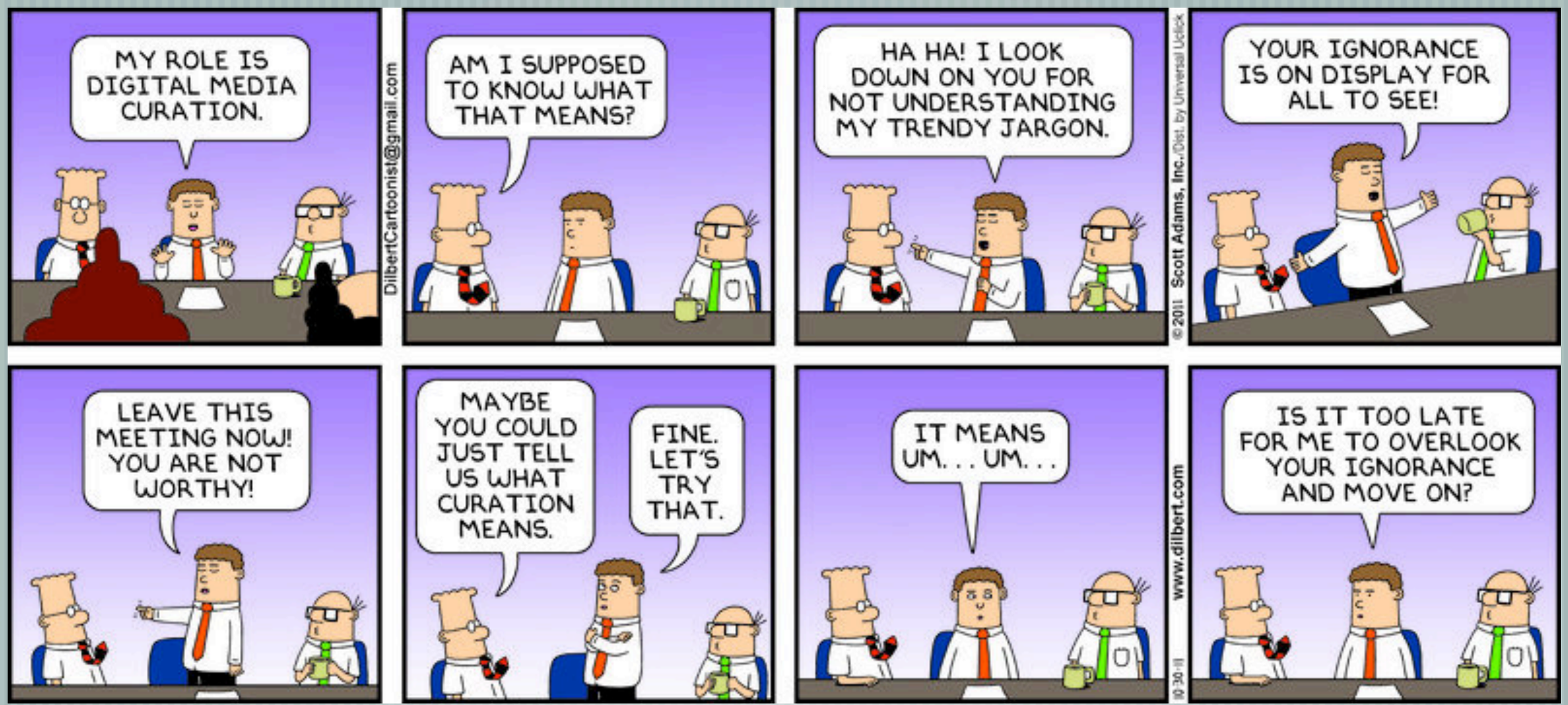
- Definitions & Concepts
  - What IS digital curation, exactly?
- Understanding the digital curation lifecycle model
- Applying the lifecycle model to research data

Wednesday, July 11, 12

- Getting our terminology straight
- Finally answering the question: “what IS digital curation?” in terms of research data
- Understanding the philosophy behind a fluid, adaptable process for preserving and presenting research data

Definitions & Concepts

What's Digital Curation?

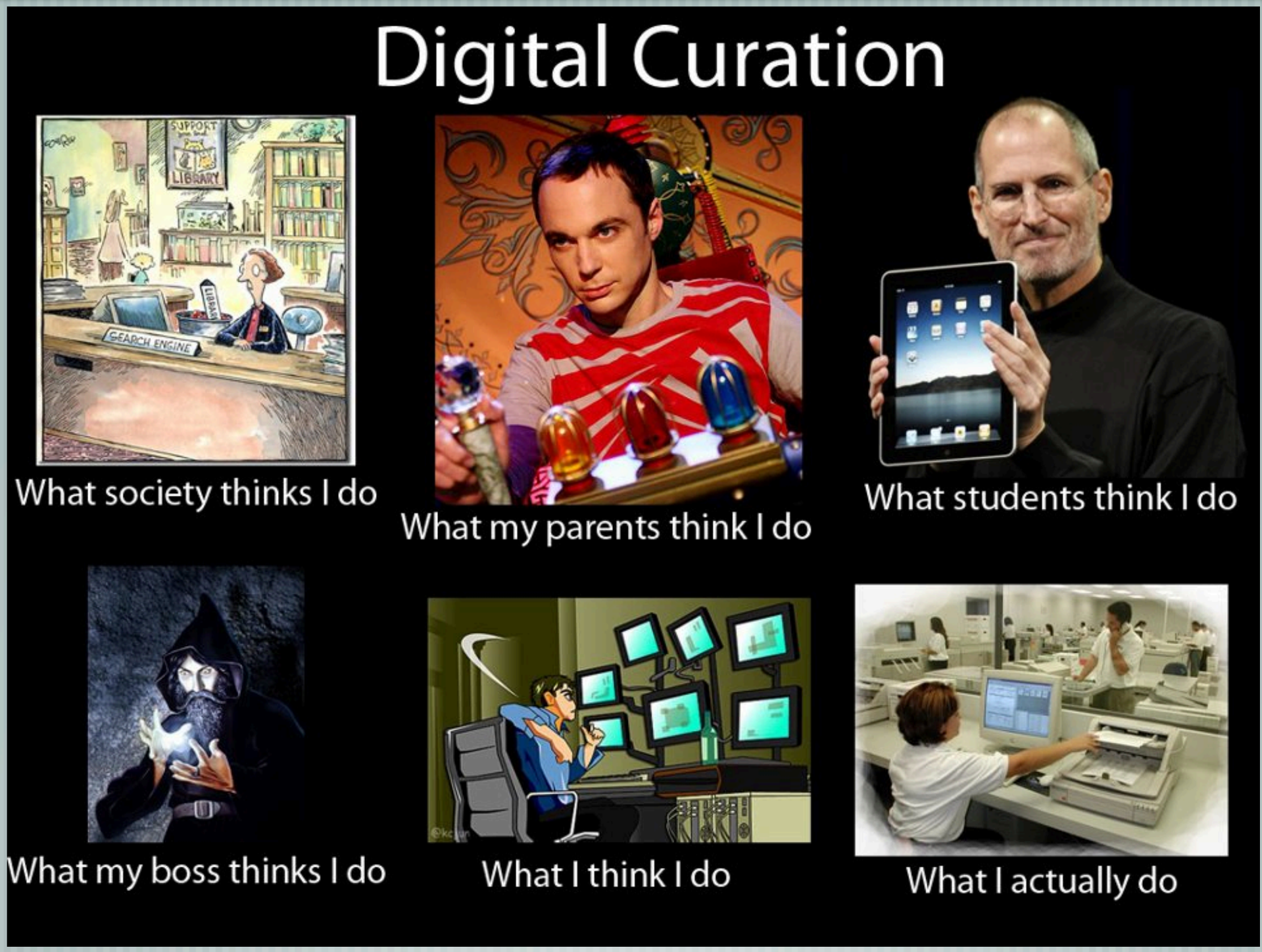


* Source: Dilbert: October 30, 2011

Wednesday, July 11, 12

• Digital Curation is an emerging field, and will continue to be a learning process. For that reason, it's not all that clear to everyone what digital curation is. You can get some confused or negative viewpoints about it.

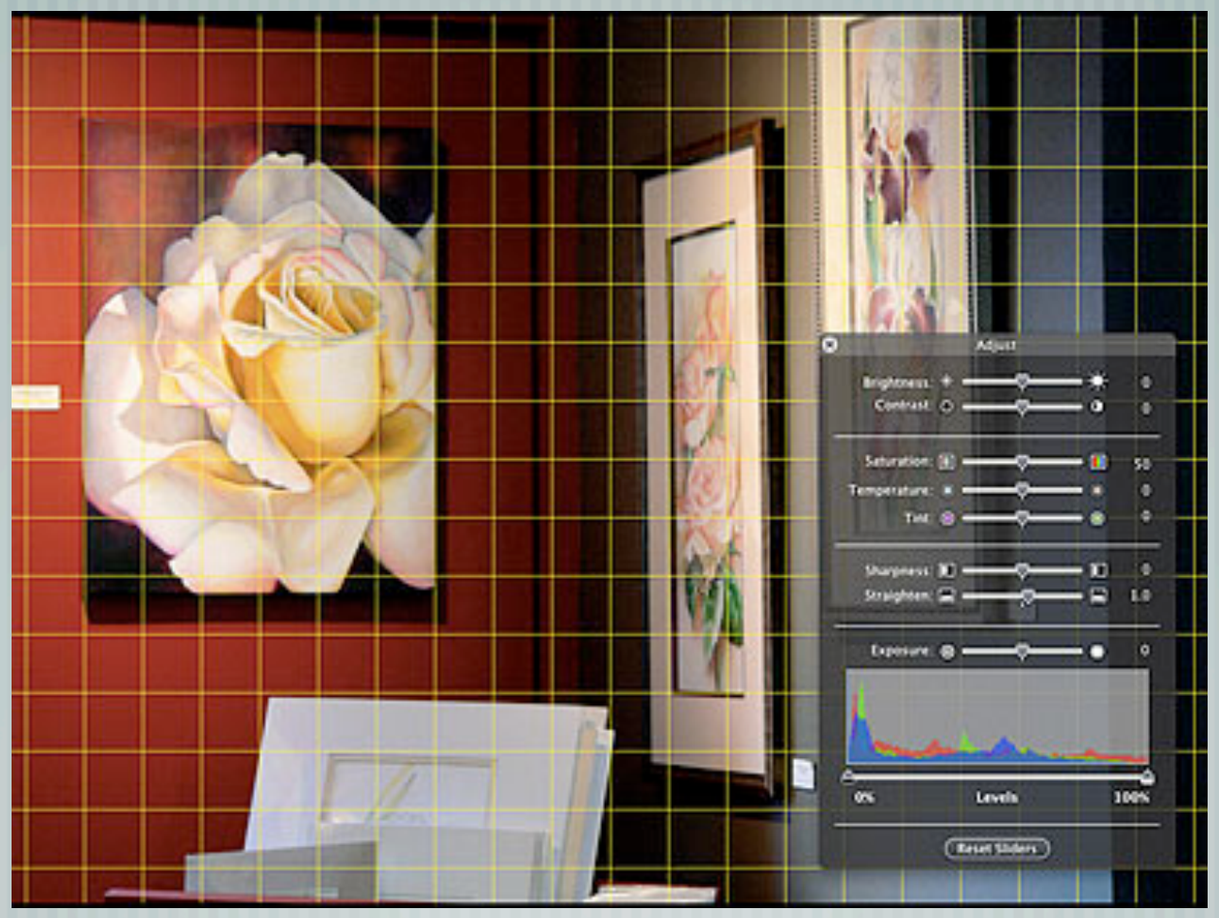
Definitions & Concepts



- There's also lots of different opinions out there about what the job entails.

Digital Curation: What is it?

"The Curation, preservation, maintenance, collection and archiving of digital assets."*



*Source: "What is Digital Curation?" Digital Curation Centre, http://www.dcc.ac.uk/about/what/

Wednesday, July 11, 12

Fortunately, the Digital Curation Centre (a UK-based JISC-Joint Information Systems Committee-funded organization) has stepped in with a clear definition on what digital curation is, and what it entails.

Definitions & Concepts

Data

- "That which is collected, observed, or created in a digital form, for purposes of analyzing to produce original research results."*
- Any related, unique information captured as part of the research process.

Dataset

- "A set of files containing both research data - usually numeric or encoded - and documentation sufficient to make the data re-usable."*

* Source: Information Services, the University of Edinburgh (2012)

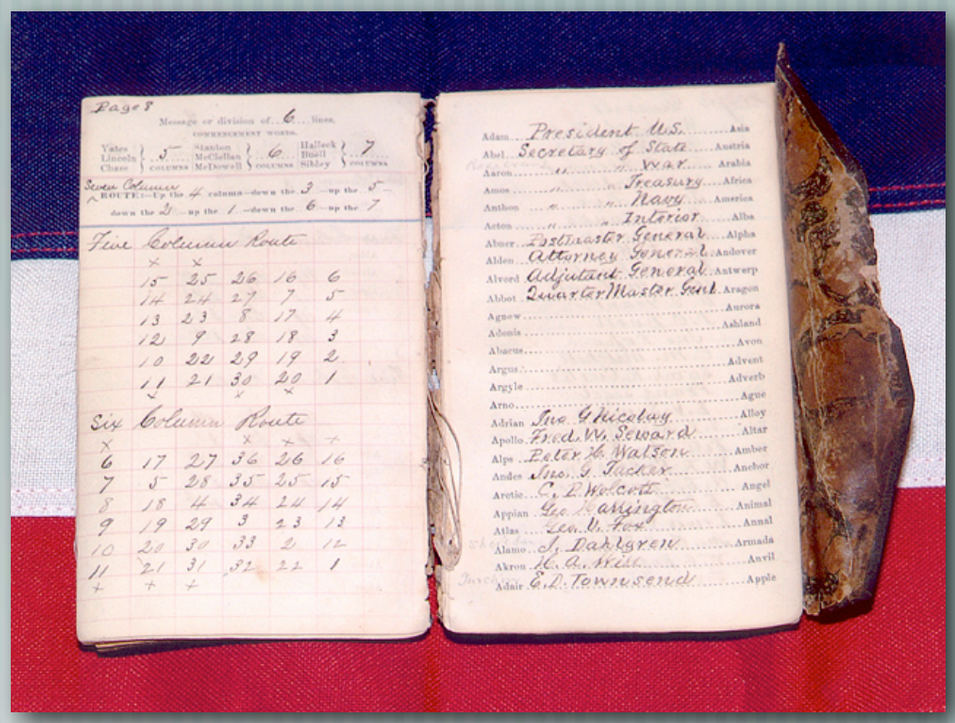
Wednesday, July 11, 12

• We're all pretty certain (hopefully) that we know what data is, but it's still important to set a specific definition so that we can frame our efforts properly. Fortunately the folks at the University of Edinburgh have done just that for us.
• Definition doesn't narrow our scope, but helps us realize the broad challenge ahead of us.

Definitions & Concepts

Documentation

- Any associated digital files which explain the research data's
- production,
- provenance,
- processing
- or interpretation



Such as a codebook, technical or methodology report, or user guide.

Wednesday, July 11, 12

- We're not just going to store data. That's not useful if people accessing the data has no idea how to interpret it.
- Documentation is necessary to make that accessibility happen.

What does a Digital Curator Do?

Acquires verifiable digital assets

- From analog sources (a "digital surrogate")

Wednesday, July 11, 12

Digital Curation involves the acquisition of digital objects. These can come from numerous sources in innumerable formats. Many objects typically come from analog sources that must be digitized. We like to assume that these are older items of historical value, in and in many cases that's true for traditional items. However, it's still possible for researchers to be capturing their initial data points using analog methods, and then convert that data into digital forms at a later step in the process.

What does a Digital Curator Do?

Acquires verifiable digital assets

- From analog sources (a "digital surrogate")
- Assets that originated digitally ("born digital")

Wednesday, July 11, 12

Others are born digital, that is, the data originated in a digital form from the start.

What does a Digital Curator Do?

Acquires verifiable digital assets

Certifies content integrity

- Minimum standards and workflow practices
- Training of staff on handling digital assets
- Quality Assurance

Wednesday, July 11, 12

Regardless of source, a digital curator must evaluate the content for file formats used, software and hardware required, the condition of the original data, and determine if choices need to be made about format stability and ease of access. The DC may also train other staff in proper procedures for handling of data objects, including their transport, conversion (if needed) and ingest. And the digital curator ensures that the resulting publicly available content is satisfactory, being handled and displayed properly.

What does a Digital Curator Do?

- **Acquires verifiable digital assets**
- **Certifies content integrity**
- **Certifies trustworthiness of the architecture**
  - Vetting of codecs and container (file) formats
  - Active role in storage decisions
  - Technical Metadata, audit trails, chain of custody

Wednesday, July 11, 12

A Digital Curator is also charged with certifying that the repository architecture is trustworthy and reliable. The Data Curator MUST know how the data is kept secure and what measures are in place to assure reliability of the platform and effective data recovery scenarios in the event of failure.

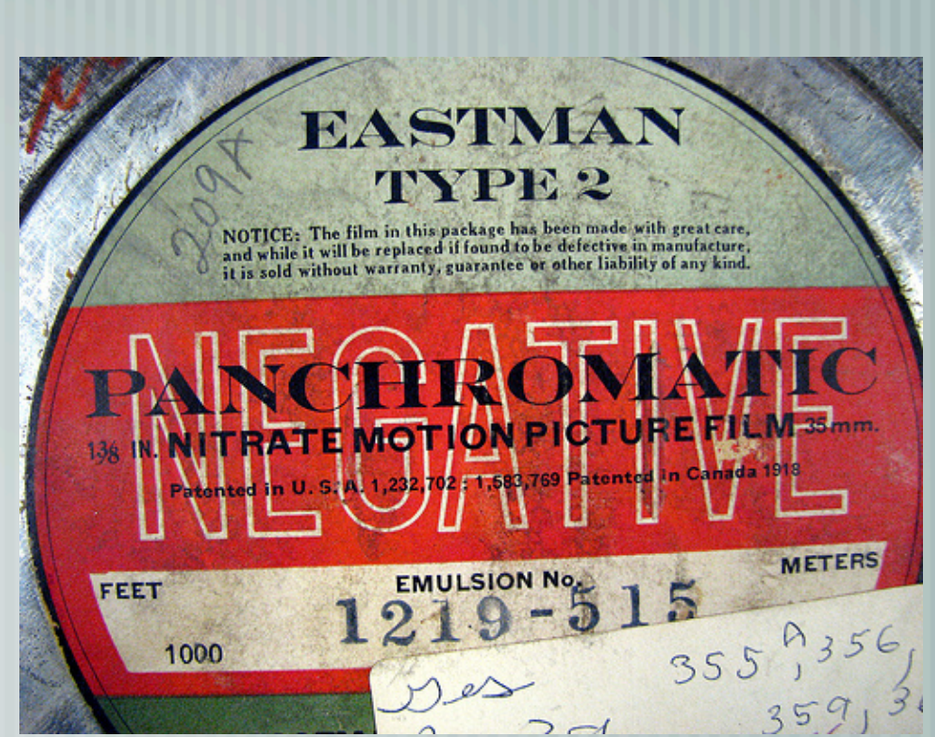
Increasing storage requirements and changes in technology mean that the DC must take an active role in researching and determining how that platform should take shape. The DC must undertstand not just the softare level, but the hardware level of storage.

Lastly, the DC is responsible for provision of technical metadata. The object most not only be stored, but it must be stored with a way of telling the history behind it; origins, the equipment and software used to crate the object, the technical specification behind the object, and all of the information an archivist needs to ensure the object's proper care and feeding. Audit trails and chains of custody also help us verify that an object was properly handled by the right people, and details what actions have been taken that may affect the object since its ingestion.

Digital Preservation Challenges

Digital Assets are easier to destroy, more readily deleted than physical objects

- Physical objects: typically stored, left behind, forgotten and "rediscovered"



Wednesday, July 11, 12

Why is all of this important?
Digital objects present a unique preservation challenge. We're all used to the "found object" model of analog artifacts. Items are stored in attics, basements and closets, sometimes because they're known to be valuable, but many times because their owners don't want to think about them and find the efforts required for disposal too onerous. They are later "rediscovered" and, assuming good condition, their historic value becomes known.

Barring disasters or poor media decisions, analog objects also tend to decay slowly over time.

Digital Preservation Challenges

Digital Assets are easier to destroy, more readily deleted than physical objects

- Digital objects: Casual collectors typically delete what they don't want when they're low on space, or see no immediate need to retain the content.

+

Wednesday, July 11, 12

Digital objects aren't so easy. We don't always see or touch our digital objects. They can be stored in containers (computers, hard drives, USB sticks, obsolete physical media) or can be completely ethereal (cloud storage). A disposal of that equipment won't reveal what important objects might be contained within. And mass destruction of data that could be important years from now is a simple keypress away.

Digital Preservation Challenges

- Digital Assets are dependent on file formats and hardware/ software platforms

Windows

webOS

redhat

solaris

chrome

android

BlackBerry

ubuntu

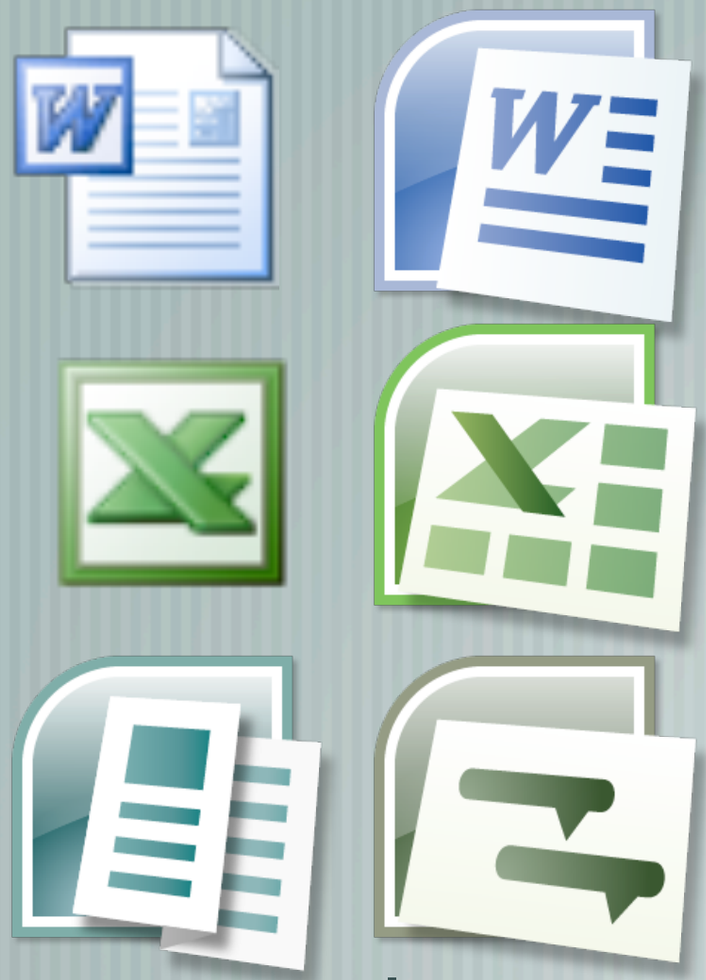
maemo™

Wednesday, July 11, 12

Digital assets are multilayered. Unlike physical, analog objects that can stand alone, digital assets rely on hardware and software to be useful. It used to be easy, with a limited number of Operating Systems and paltforms. But the space has grown far more diverse recently, and now there are many platforms out there where research data is created, recorded and stored, and we must make ourselves familiar with them.

Digital Preservation Challenges

Digital Assets are dependent on file formats and hardware/ software platforms



Born Digital Documents

Still Images

Sound files

- At least 27 common file formats
- At least 90 common codecs

Moving Images

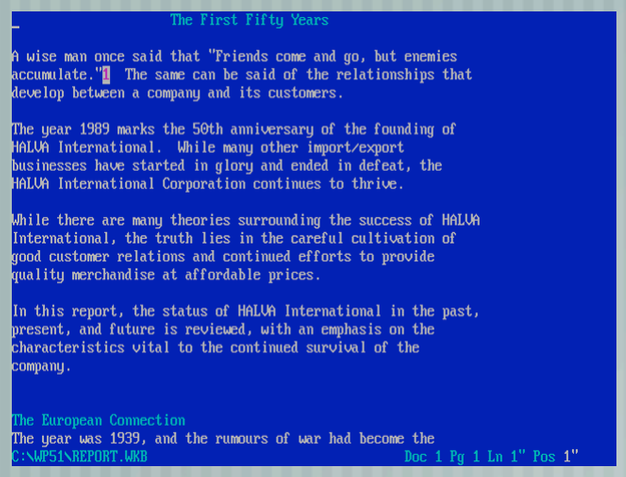
- At least 58 common containers/codecs
- + Audio tracks (27 formats/90 codecs)

Wednesday, July 11, 12

The Operating System is but one layer. Data is created and stored using countless software packages producing countless file types and formats, which evolve and change over time. Here, I'm only mentioning the formats we're familiar with. Data sets can stored in the above formats, or in plenty of other formats we haven't even heard of yet.

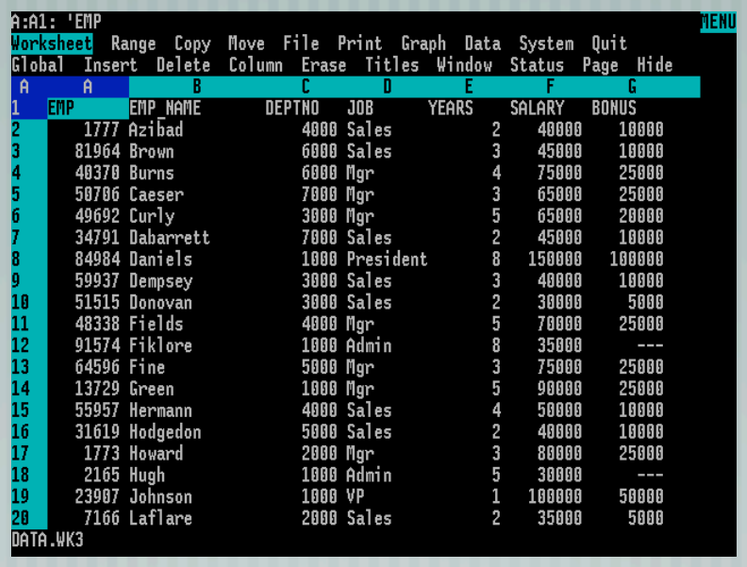
The Threat of a "Digital Dark Age"

Digital Assets are vulnerable to Format Obsolescence



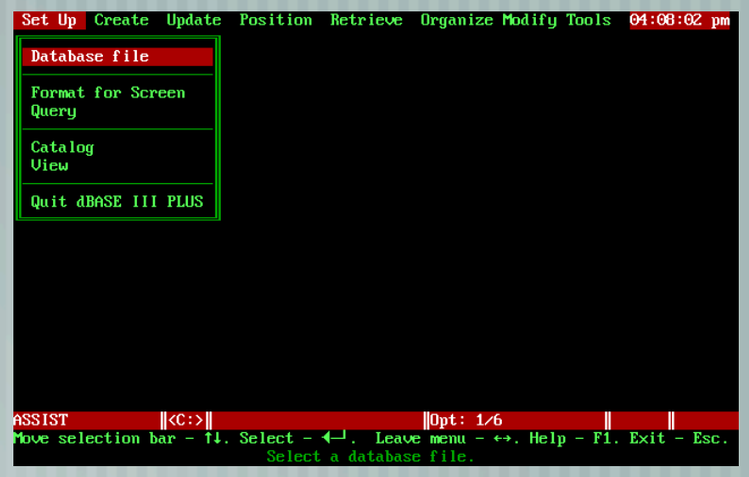
Wordperfect

- 1979



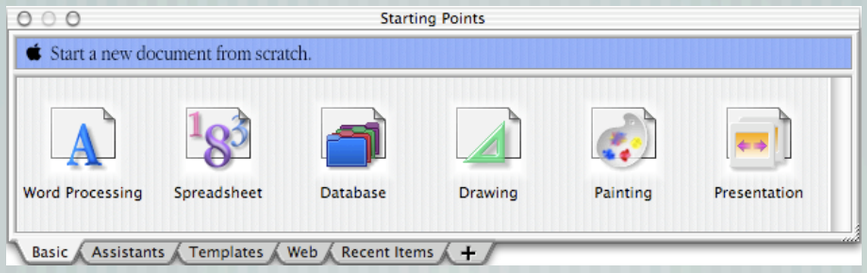
Lotus 1-2-3

- 1978



Dbase

- 1978



Appleworks

- 1984

Wednesday, July 11, 12

This ever-evolving, ever-changing landscape of files and formats means that format obsolesce is inevitable. Old software evolves into newer versions. If they don't evolve, they give way to new alternatives that replace the old. File formats become deprecated, requiring migration to new formats over time. Even basic text files are no longer the same as they were decades ago.

The Threat of a "Digital Dark Age"

Digital Assets are vulnerable to Format Obsolescence

Tape

Floppy Disk

Zip Disk

IBM "Demi-disk"

Video Floppy

Wednesday, July 11, 12

The software isn't the only problem. Hardware and media evolves and changes over time, as well.

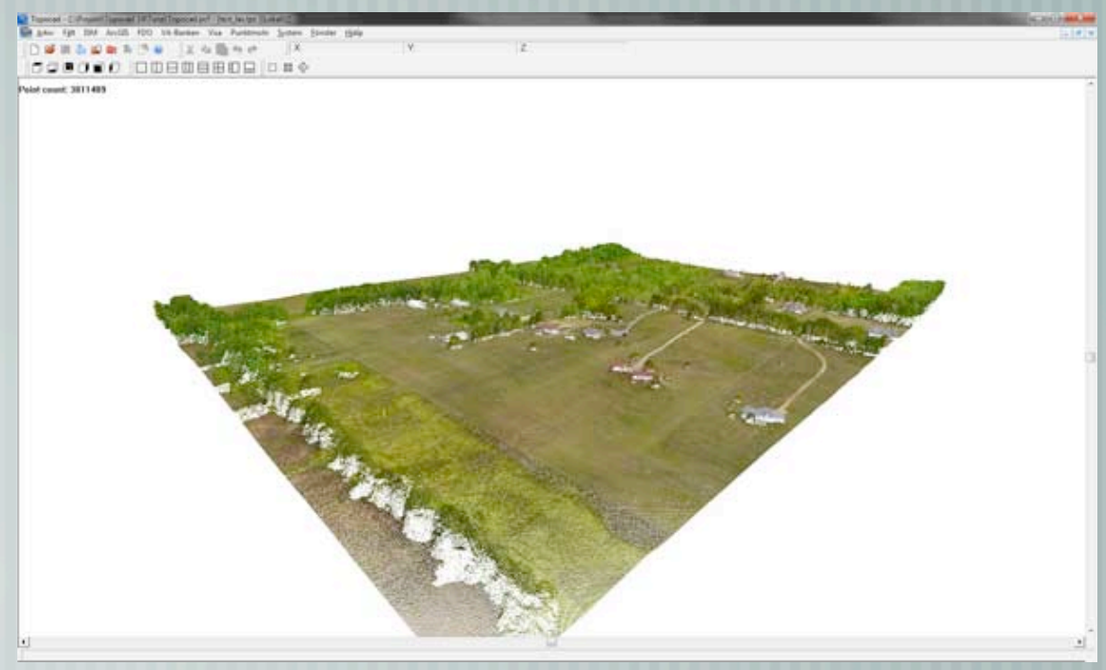
Additional challenges: Datasets & File Formats

- Can be anything.
  - Already known and established formats, OR
  - Totally new formats: SUR, CSFASTA...

```
;LCBO - Prolactin precursor - Bovine
; a sample sequence in FASTA format
MDSKGSSQKGSRLLLLLVVSNLLLCQGVVSTPVCPNGPGNCQVSLRDLFDRAVMVSHYIHDLSS
EMFNEFDKRYAQGKGFITMALNSCHTSSLPTPEDKEQAQQTHHEVLMSLILGLLRSWNDPLYHL
VTEVRGMKGAPDAILSRAIEIEEENKRLLEGMEMIFGQVIPGAKETEPYPVWSGLPSLQTKDED
ARYSAFYNLLHCLRRDSSKIDTYLKLLNCRIIYNNNC*
```

```
>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK*
```

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```



Wednesday, July 11, 12

The responsibility falls upon us to step out of that comfort zone, learn what we need to learn about these new file types and how they are created and handled, and craft a preservation plan that does right by those sets of research data. We will not always have the answers immediately, but we will need to learn as we go.

Additional challenges: Datasets & File Formats

Most traditional file types are "known quantities."

- Predictable use cases
- Rigid standards
- Built-in familiarity

Wednesday, July 11, 12

For the most part, we tend to know what we're doing when presented with a typical, traditional file type. We know what people do with MS Office documents or PDFs. We know how to display still images, or play sound files and movies. Most of these formats are pretty self-explanatory, and their preservation needs are predictable and can be applied pretty uniformly. We are in our comfort zone with these file types.

Additional challenges: Datasets & File Formats

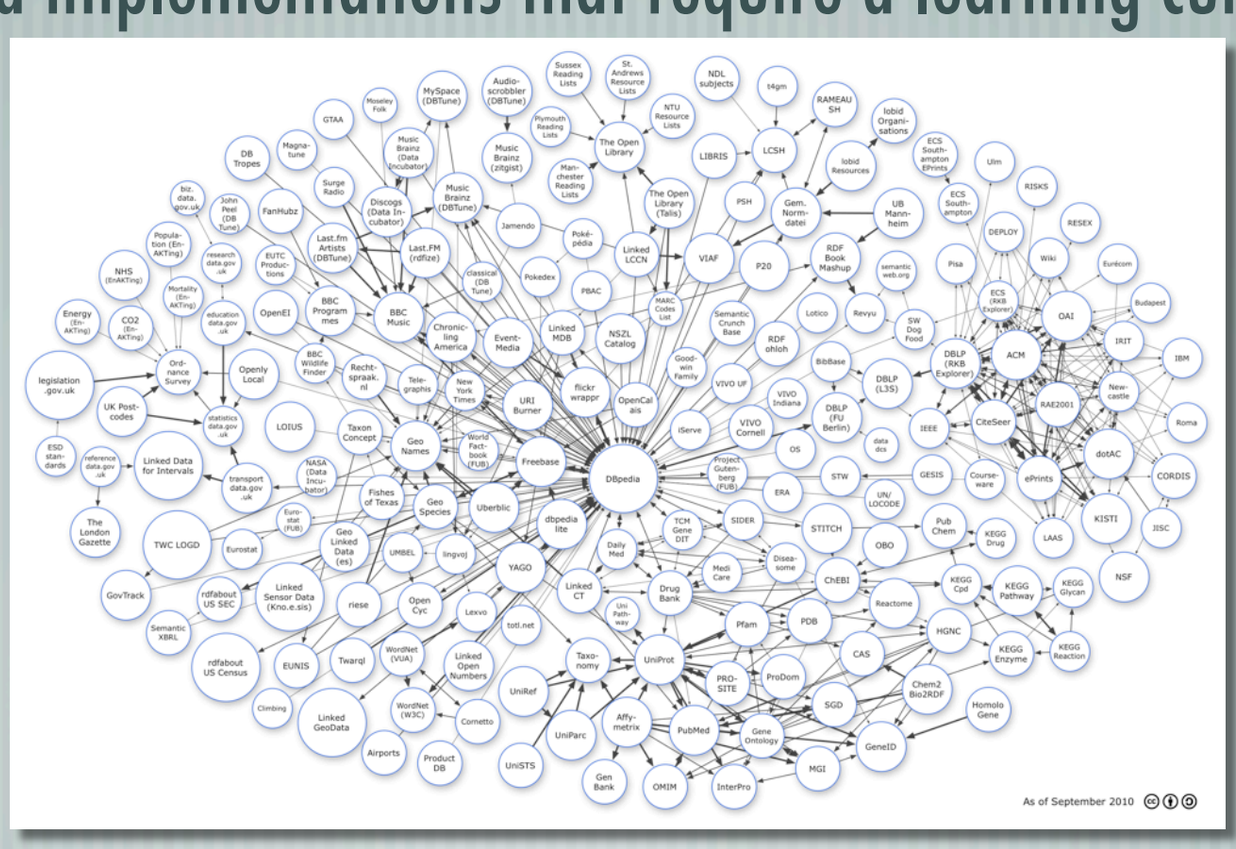
- Some datasets are stored merely as common file formats used for the purpose of data gathering, e.g.
  - MS Excel spreadsheets with data points, PDF files with written content.
  - Still images, sound, or moving images captured as part of research.

Wednesday, July 11, 12

Some research data will be amenable and comformable to what we already know. Some research data will come in familair formats: Spreadsheets or CSVs, PDFs, text files, moving images and sound.

Additional challenges: Datasets & File Formats

- **Research Data will take out us out of our comfort zone**
  - Unique, non-traditional, obscure data types
  - Traditional files used in non-traditional ways
  - Usages and implementations that require a learning curve to the uninitiated.

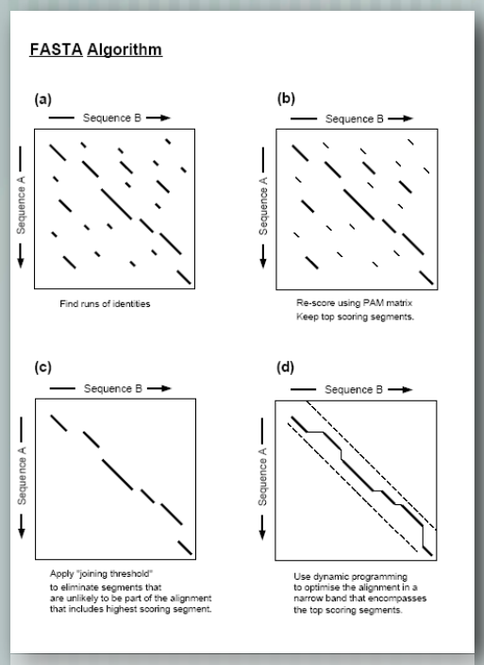


Wednesday, July 11, 12

But datasets can be unpredictable, threatening to take us out of that comfort zone. They will represent unique ways of handling information that we haven't encountered up to now, but form a necessary part of a researcher's workflow.

Additional challenges: Datasets & File Formats

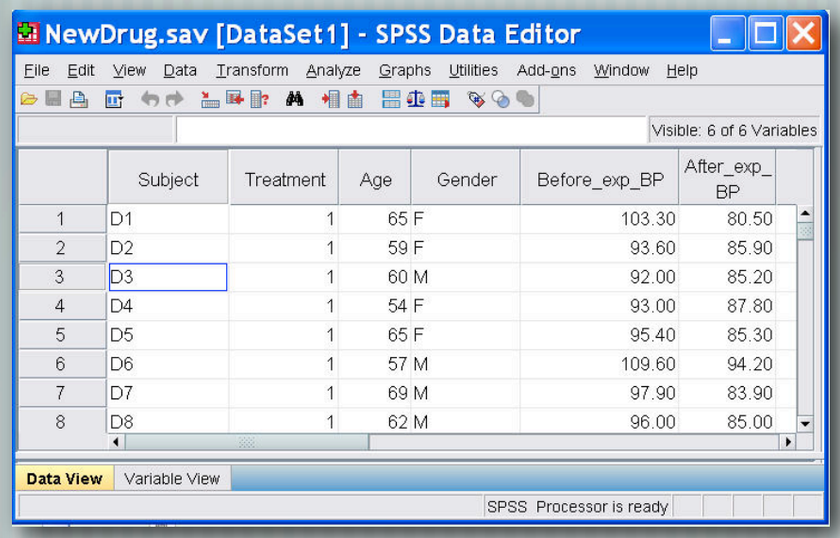
- Other dataset file formats can be extensions of existing file types that are re-purposed. Can be human-read, or interpreted with additional, special-purpose software.
  - e.g. Repurposed UTF-8 (text file) to create a FASTA sequence.

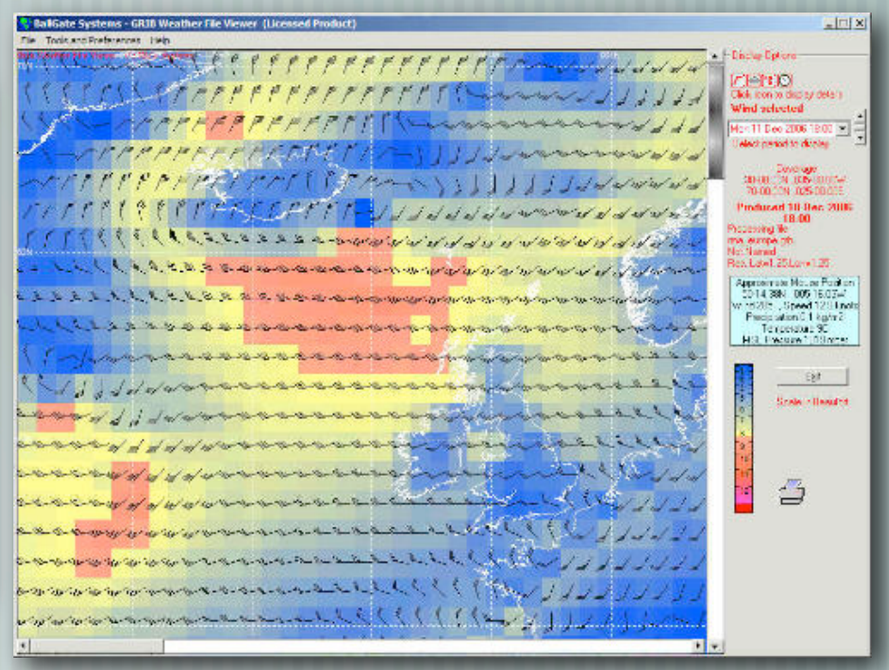



```
;LCBO - Prolactin precursor - Bovine
; a sample sequence in FASTA format
MDSKGSSQKGSRLLLLLVVSNLLLCQGVVSTPVCPNGPGNCQVSLRDLFDRAVMVSHYIHDLSS
EMFNEFDKRYAQGKGFITMALNSCHTSSLPTPEDKEQAQQTHHEVLMSLILGLLRSWNDPLYHL
VTEVRGMKGAPDAILSRAIEIEEENKRLLEGMEMIFGQVIPGAKETEPYPVWSGLPSLQTKDED
ARYSAFYNLLHCLRRDSSKIDTYLKLLNCRIIYNNNC*

>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK*

>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

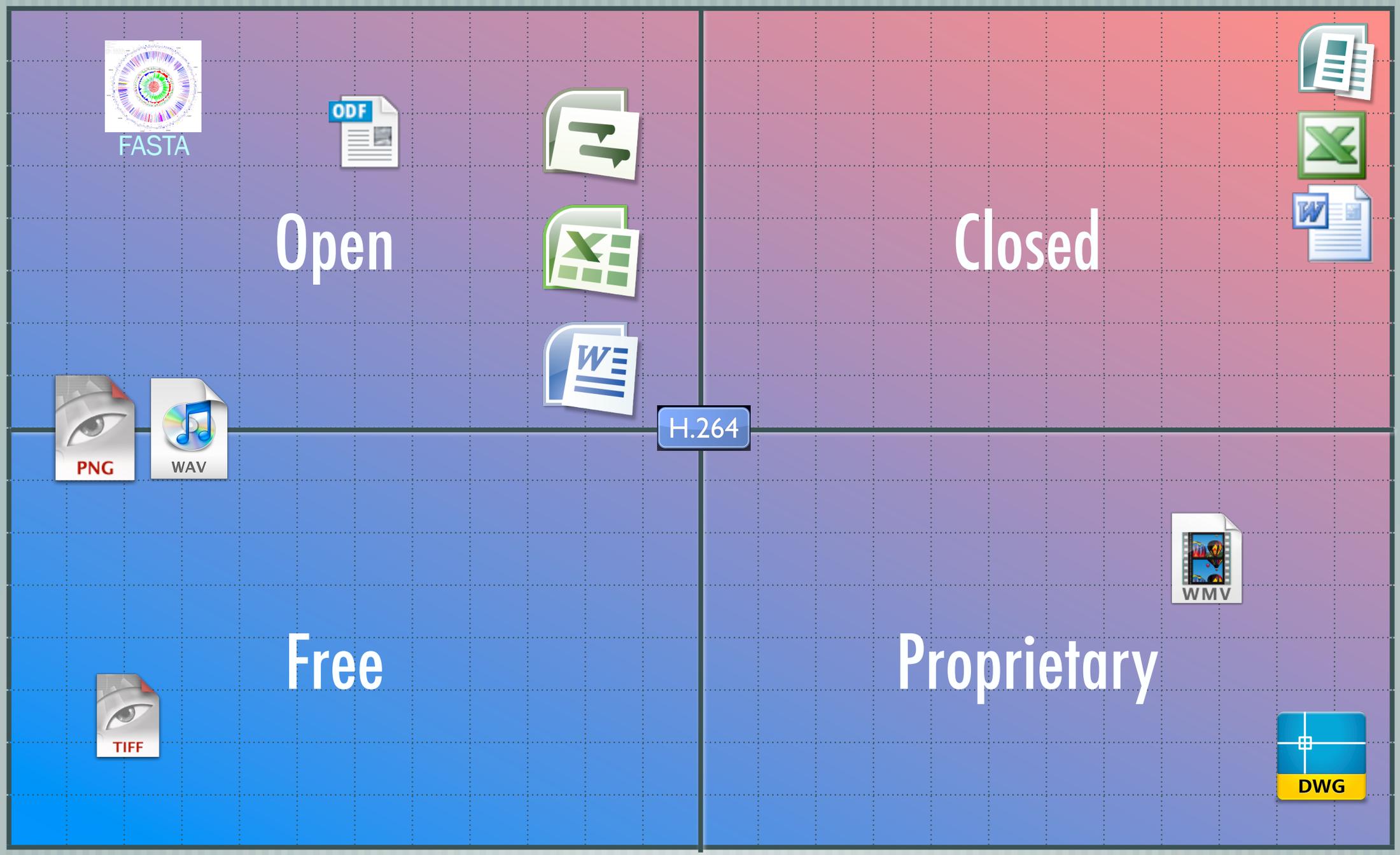
Additional challenges: Datasets & File Formats

- Finally, datasets can be completely proprietary, custom and closed, requiring specialized hardware/software to access or interpret.
  - e.g. GRIB, SUR, DWG, SPSS... too many to list!





File Formats: Open/Closed vs. Free/Proprietary



Wednesday, July 11, 12

The makeup of a file format isn't just dependent on the software that creates it, but under what principles the structure of that format is released under. Some software and files types are free and open; there's no patent of commercial restriction to their use. Other file types are parts of closed systems; restricted by their inventors, requiring commercial software to access the data contained inside. Some free file types aren't popular or well supported, putting their longevity at risk. The same can be true of tightly-restricted formats that aren't adopted on a wide scale. These factors must be taken into account when evaluating their preservability.

Lots of issues here. How do we deal with these challenges across a disapora of varied and unpredictable situations?

The Digital Curation Lifecycle

- A multitiered, continuous process where digital objects of any type are evaluated, preserved, maintained, verified, and re-evaluated.
- Iterative: the cycle doesn't end with one go-round.
- A useful exercise for known and as-yet-unknown file types and formats.

Wednesday, July 11, 12

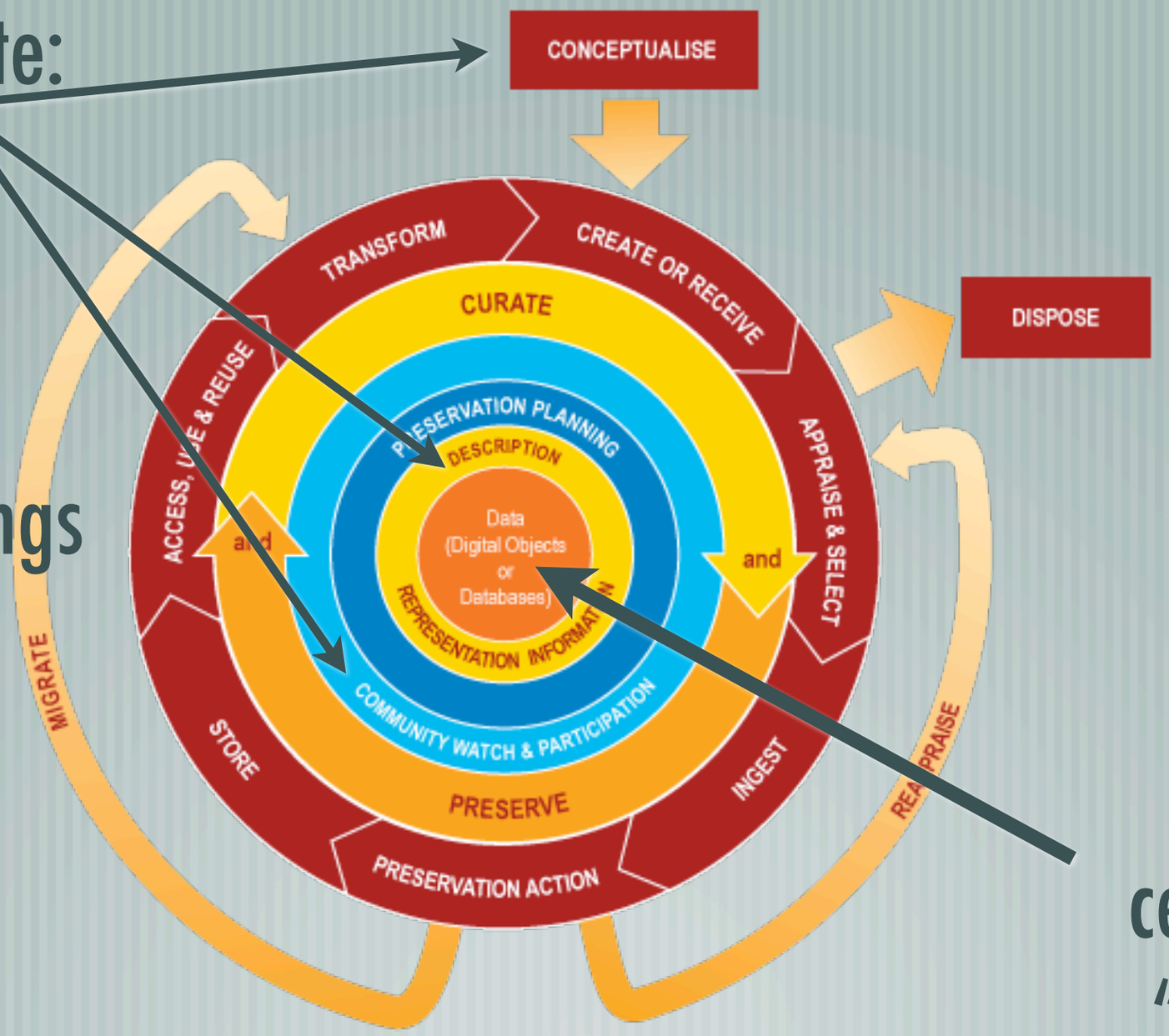
Fortunately, there is an established process for addressing all of these issues. It's called the Digital Curation lifecycle.

Here's what it looks like...

The Digital Curation Lifecycle

Collaborate:

- Plan
- Describe
- Evaluate
- Learn meanings



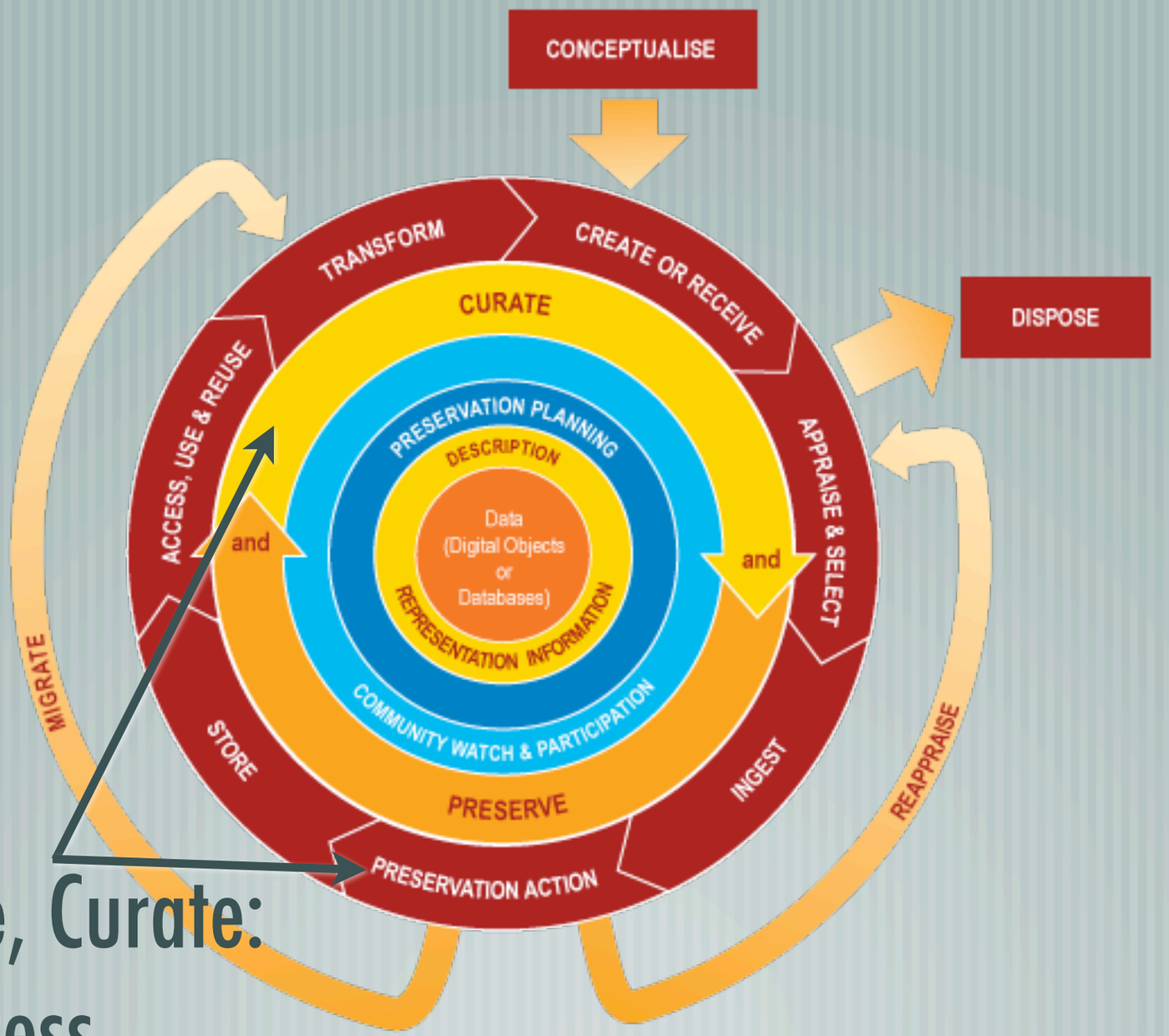
Data: The center of our "universe."

* Source: Digital Curation Centre, http://www.dcc.ac.uk/resources/curation-lifecycle-model

Wednesday, July 11, 12

The model starts with the data, which is prominent at the center of this model. But just as important is collaboration as part of the planning and learning process. It is a common misconception that data is created or captured by a researcher, and then the proper way to proceed is to simply pass it on to someone else to curate. In fact, much of the most crucial information required for effective long-term curation and reuse must be captured at the conceptualization and collection stages. If the researcher isn't actively curating their data, then they MUST communicate and collaborate with those who are. This stage is where the data files are evaluated, formats checked and researched, a data model is constructed, and the RUresearch data team obtains the infromation they need to move forward with a preservation plan.

The Digital Curation Lifecycle



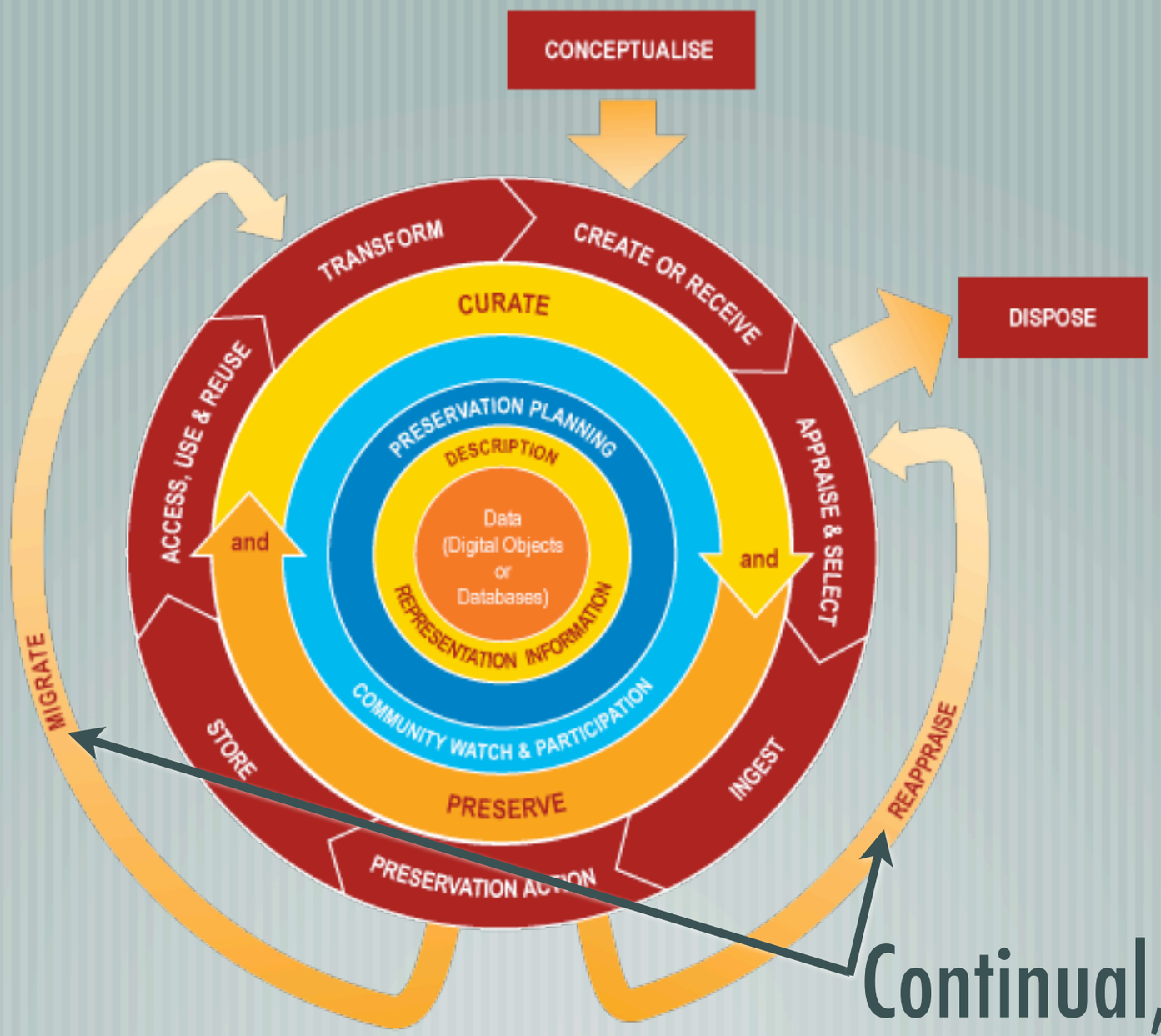
Ingest, Preserve, Curate: The RUcore Process

* Source: Digital Curation Centre, http://www.dcc.ac.uk/resources/curation-lifecycle-model

Wednesday, July 11, 12

Only after the communicative and collaborate process can the cataloging, preparation, planning and ingest of the data objects take place. The RUresearch team doesn't need to be perfect experts in the field of study pertinent to the dataset, but they need to be familiar with what the data represents and how it is to be used. Communication with the researcher doesn't end though: they are a partner to this process and their feedback is required as we progress with the RUcore process.

The Digital Curation Lifecycle



Continual, iterative post-ingest actions

* Source: Digital Curation Centre, http://www.dcc.ac.uk/resources/curation-lifecycle-model

Wednesday, July 11, 12

Ingesting the dataset and making sure it displays properly in search results isn't the end of the story. As with every item in RUcore, the lifecycle continues, and is a never-ending process. As data ages on the storage server, it must be checked for signs of corruption or deterioritation. File formats must continue to be evaluated to see if they continue to be supported, or if they have been superseded. Migration to new formats must occur as needs warrant. And we must periodically reassess how we manage, store and present our collections and determine how things can be done better with the technology at hand as it advances.

Solutions: A "controlled chaos"

Evaluate the data, the research project, and the researcher's needs. Creation of a descriptive, comprehensive data model for the project is key.

- Take stock of Software, Systems, measuring/lab equipment, and recording apparatus.
- Often, we must accept that de facto industry/ research standards become de facto preservation standards.

Solutions: A "controlled chaos"

- Establish a format guide and handling procedures. Evaluate the veracity and longevity of the data format. Check competitors, alternatives, and potential successor formats. Publish, share and use the findings.

- Determine methods of access. How are users expected to access and view the data?
  - Software/hardware requirements
  - View online? Use a plug-in? Download and use third party software?

Solutions: A "controlled chaos"

- Do No Harm to digital assets
  - Preservation masters, derivatives when needed
  - Content modification must be done with extreme care
  - Any changes must be traceable, auditable, reversible

Solutions: A "controlled chaos"

- Prepare for the inevitable: format migrations
  - Periodically re-assess the relevant format
  - Migrate to new formats when the old is obsolete
  - Maintain accessibility while ensuring data integrity

Case Study: Cranberry Genome

- Philip E. Marucci Center for Blueberry and Cranberry
  - Cranberry (Vaccinium macrocarpon) clone CNJ99-125-1 genome, July 2009 extraction
  - FASTA and QUAL files created from SOLiD, forward and reverse genome sequence
  - Dataset: 4 large files, gzipped, totaling 40GB of data, compressed.

Case Study: Cranberry Genome

- Triage
  - What is FASTA/QUAL?
  - Determine lifecycle events
  - Determine access/presentation scenarios

Case Study: Cranberry Genome

Triage

What is FASTA/QUAL?

"...a text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes. The format also allows for sequence names and comments to precede the sequences. The format originates from the FASTA software package, but has now become a standard in the field of bioinformatics.

The simplicity of FASTA makes it easy to manipulate and parse sequences using text-processing tools and scripting languages like Python, Ruby, and Perl."*

*Source: NIH

Case Study: Cranberry Genome

- Triage
  - What is FASTA/QUAL?
  - Determine lifecycle events
    - Origins
      - Researchers
      - Grant Awards
    - Purpose
    - Software/Equipment

*Source: NIH

Case Study: Cranberry Genome

- Triage
  - What is FASTA/QUAL?
  - Determine lifecycle events

Data Life Cycle Event(s)

Type: Gene sequencing

Label: Vaccinium macrocarpon clone CNJ99-125-1, a fifth-generation inbred clone derived from self-pollination of the cultivar Ben Lear.

Detail: Sequencing took place at the Waksman Genomics Core Facility (contact David Sidote) using the Applied Biosystems SOLiD 3 Plus. Sequencing methodology was Mate-paired library, 1.5-2 kilobase pair (kb) fragments.

Service provider: Waksman Genomics Core Facility

Name: Applied Biosystems SOLiD 3 Plus System

Reference: http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_072050.pdf

*Source: NIH

Case Study: Cranberry Genome

- Triage
  - What is FASTA/QUAL?
  - Determine lifecycle events
  - Determine access/presentation scenarios
    - Web-based presentation?
    - or present the file as-is?

Questions?

Additional resources:

Digital Curation Center:

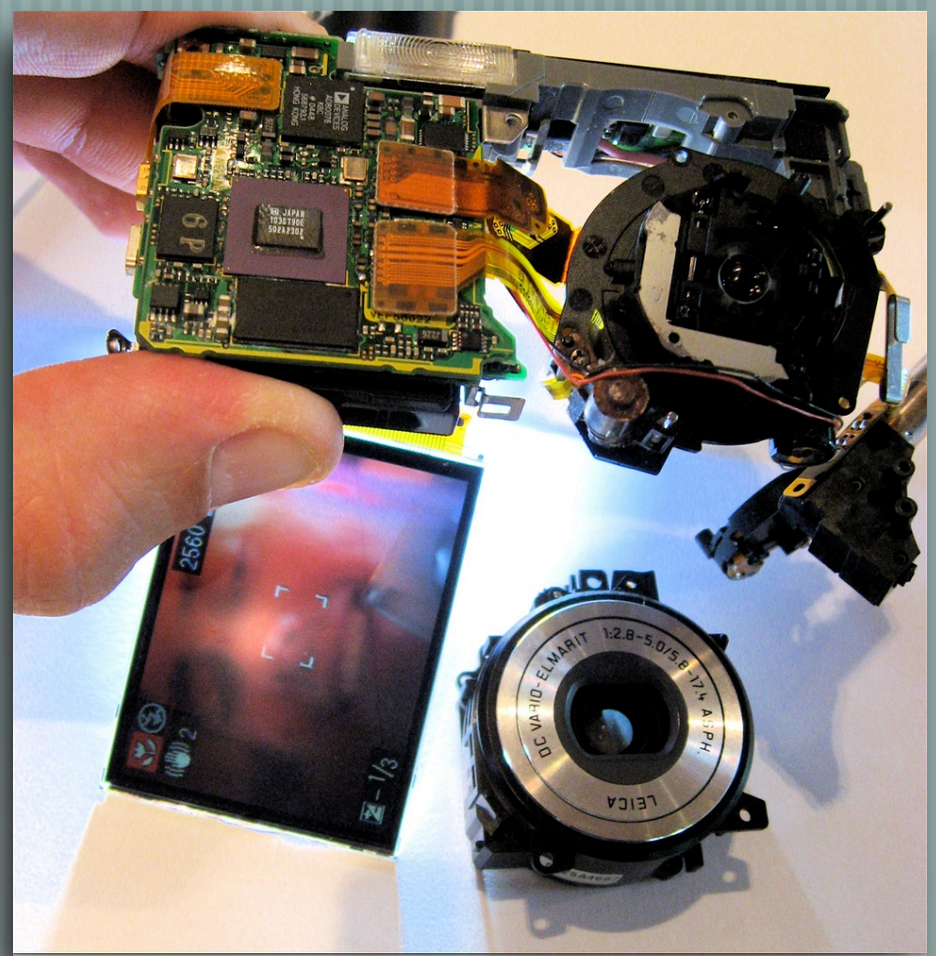
- www.dcc.ac.uk

Australian National Data Service:

- www.ands.org.au

Blog: From Page2Pixel

- page2pixel.org



WORLD BACKUP DAY 2012

MARCH 31ST

DON'T BE AN APRIL FOOL.
BACKUP YOUR FILES.
CHECK YOUR RESTORES.

Remember to ensure your files are backed up on **March 31st**.

Backup your memories and financial information and check your old backup restores.

www.worldbackupday.com