LIFE: Costing the Digital Preservation Lifecycle

Paul Wheatley

Digital Preservation Manager

The British Library

Summary

- Aims
- The LIFE and LIFE2 Projects
- The LIFE Model
- Case studies
- The Generic Preservation Model
- Looking ahead: LIFE3

LIFE

Digital Preservation. A question of....

...how?

...how much?

LIFE

Objectives

- Better understanding of the digital lifecycle
- An ability to plan and prepare for digital preservation activities
- Evaluate and improve our efforts
- Compare analogue and digital

LIFE projects overview

- Collaboration between UCL and the British Library
- Co-funded by Joint Information Systems Committee (JISC)
- The LIFE Project:
  - 1 year project
  - Completed in Spring 2006
- The LIFE2 Project:
  - 1.5 year project
  - Began in Spring 2007

LIFE = **L**ife cycle **I**nformation **F**or **E**-literature

Overview of project focus

- **The LIFE Project:**
  - Aim: to explore a lifecycle approach to costing the preservation of digital materials
  - Developed:
    - A model of the digital preservation lifecycle
    - A predictive tool for estimating preservation costs
    - 3 case studies, examining real life digital lifecycles
- **The LIFE2 Project:**
  - Aim: to evaluate, refine and further develop the techniques developed in phase one of LIFE
  - Key elements:
    - Review by external economics expert
    - A refined lifecycle model and detailed costing methodology
    - 5 new lifecycle case studies

Introducing the LIFE Model v1.1

Introducing the LIFE Model v1.1

Lifecycle Stage	Creation or Purchase	Acquisition	Ingest	Metadata Creation	Bit-stream Preservation	Content Preservation	Access
Lifecycle Elements	....	Selection	Quality Assurance	Re-use Existing Metadata	Repository Admin	Preservation Watch	Access Provision
	....	Submission Agreement	Deposit	Metadata Creation	Storage Provision	Preservation Planning	Access Control
	....	IPR & Licensing	Holdings Update	Metadata Extraction	Refreshment	Preservation Action	User Support
	....	Ordering & Invoicing	Reference Linking		Backup	Re-ingest	
		Obtaining			Inspection		
		Check-in					

Publishing the Model: www.life.ac.uk

- Life Model V1.1 will be released October 19th

Publishing the Model: www.life.ac.uk

- Life Model V1.1 will be released October 19th
- Detailed definitions, and sub element descriptions

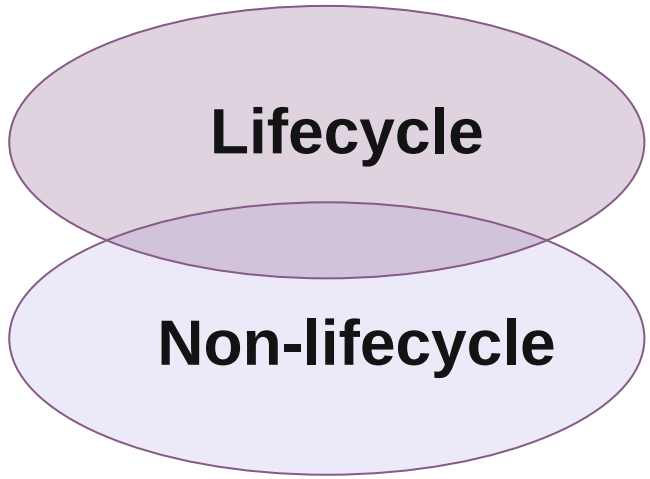
1.Quality Assurance

Quality Assurance is the process of examining digital objects and ensuring they are of a sufficient or expected level of quality. If the assessed quality level is not sufficient, a mitigation strategy might have to be applied. This might include applying fixes, re-acquiring objects or recording metadata describing the details of the quality issues encountered. QA typically includes the process of checking the materials for viruses, and taking appropriate action to clean virus tainted objects.

Suggested Sub-elements	Explanation / notes
QA Policy (policy/procedure)	Description of quality requirements and required mitigation actions should quality requirements not be met. Policy for sampling of objects for QA (if appropriate)
QA Characterisation (action)	Characterisation of the digital object. Identification of file format, and assessment of whether the object is valid, well formed, and/or renders correctly with current access software
Content Examination (action)	Assessment of whether the content of the digital object is of an expected, agreed or sufficient level of quality. Typically, a manual process on a sample of the ingested objects
Mitigation (action)	Action to mitigate quality issues (might include virus cleaning or re-ordering or obtaining the digital object)
QA Metadata (metadata)	Record QA metadata

Publishing the Model: www.life.ac.uk

- Life Model V1.1 will be released October 19th
- Detailed definitions, and sub element descriptions
- Scoping lifecycle costs



LIFE

Publishing the Model: www.life.ac.uk

- Life Model V1.1 will be released October 19th
- Detailed definitions, and sub element descriptions
- Scoping lifecycle costs

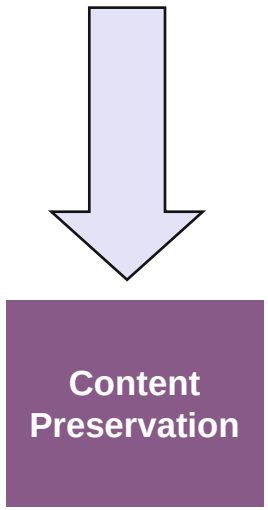
Feedback to life@bl.uk

Content Preservation

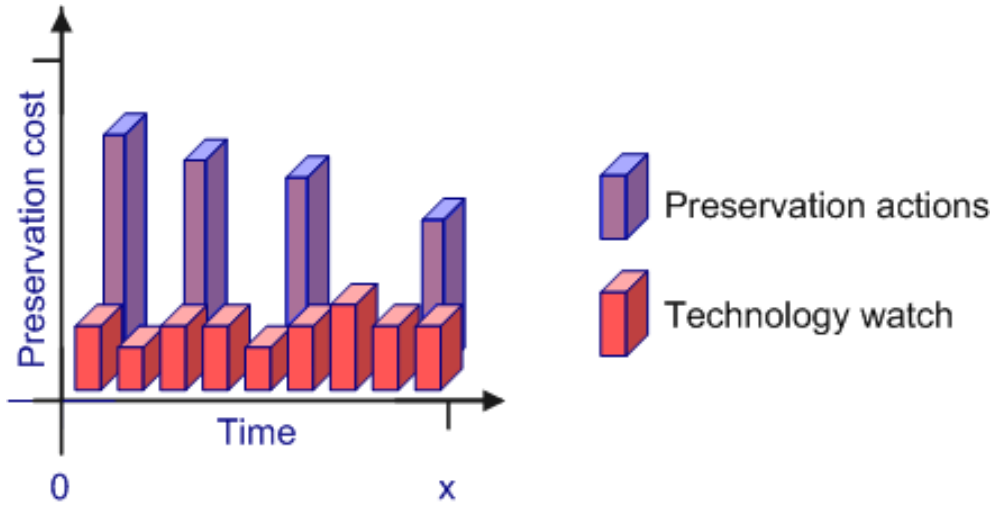
Lifecycle Stage	Creation or Purchase	Acquisition	Ingest	Metadata Creation	Bit-stream Preservation	Content Preservation	Access
Lifecycle Elements	....	Selection	Quality Assurance	Re-use Existing Metadata	Repository Admin	Preservation Watch	Access Provision
	....	Submission Agreement	Deposit	Metadata Creation	Storage Provision	Preservation Planning	Access Control
	....	IPR & Licensing	Holdings Update	Metadata Extraction	Refreshment	Preservation Action	User Support
	....	Ordering & Invoicing	Reference Linking		Backup	Re-ingest	
		Obtaining			Inspection		
		Check-in					

Introducing the Generic LIFE Preservation Model

Predictive cost model



- No data on which to cost this stage
- Identify key activities
- Model the trends



Developing the Generic LIFE Preservation Model

- Identify the main preservation activities and influencing factors
- Include key figures as editable inputs to the model
- Review and refine the model:
  - Independently model trends and map to the model
  - Review by BL Architecture and Preservation teams
  - Assess results on real content, using the LIFE1 case studies

LIFE

The Generic LIFE Preservation Model

Preservation = t * TEW + (t / ULE + PON) * (CRS + UME + PPA + QAA)

Expansion of calculated components:

- ULE – Unaided Life Expectancy of a Format = BLE + 0.1*t
- CRS – Cost of new rendering solution = (1 - PTA) * TDC * FCX + PTA * COA
- PPA – Performing preservation action = PON * (SCM + n * HVM)
- QAA – Quality Assurance = n * BCT * FCX
- PTA – Proportion of Tool Availability = STA(1-t/20)+ETA(t/20)

Expansion of scaling components:

- PON – Proportion of normalisation = 0.4
- FCX - Format complexity (e.g. JPEG = 0.2, WMF = 0.4, PDF = 0.6, Word = 0.8)

Expansion of cost component inputs:

- HVM – High volume migration cost per object = £0.05
- BCT – Base cost of testing a preservation action per object = £0.17
- UME – Update Metadata = 2 metadata officer weeks @ £30k annual salary = £1250
- TDC – Tool development cost = 24 programmer months @ £30k annual salary - £60000
- COA – Cost of available tool = £1500
- TEW - Technology Watch = 1 metadata officer week @ £30k annual salary = £625
- BLE - Base life expectancy = 8 (years)
- STA – Starting tool availability = 0.5
- ETA – Ending tool availability = 0.9
- SCM – Setup cost of migration = £340

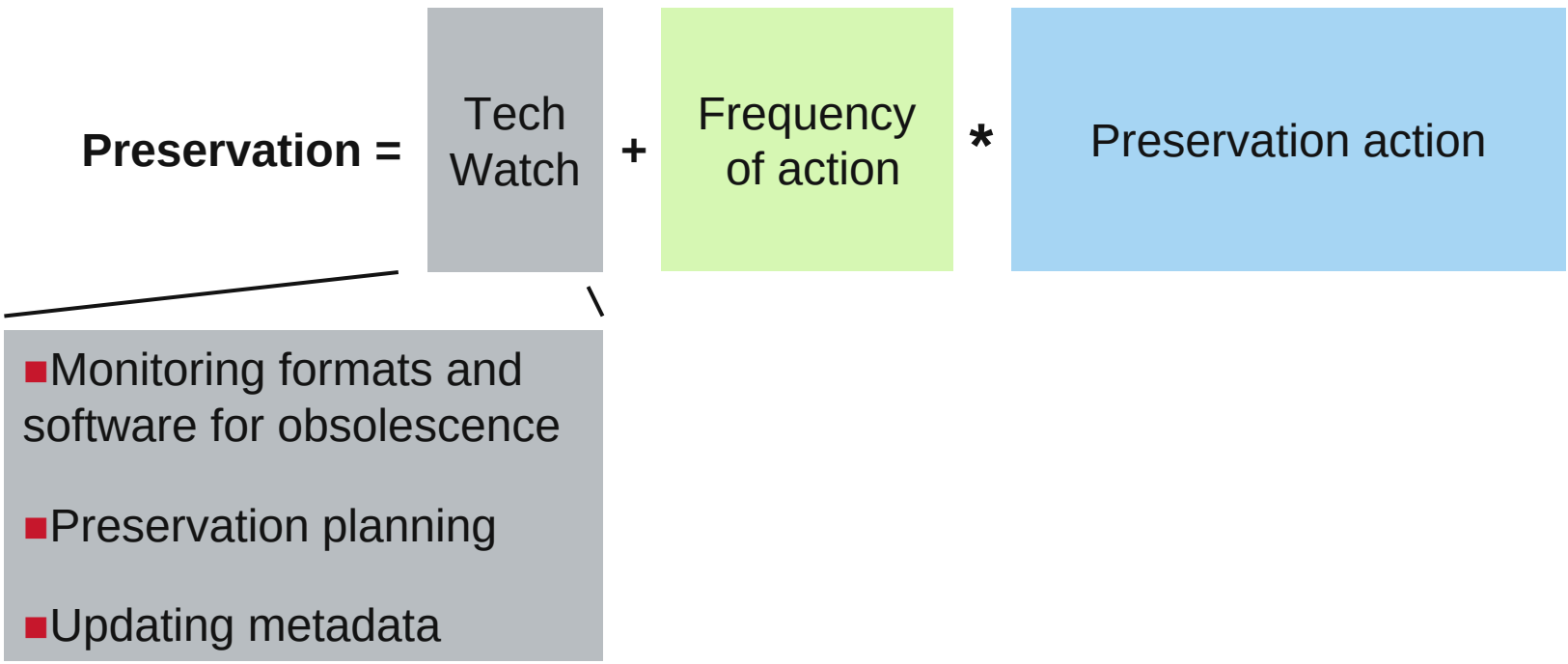
LIFE

The Generic LIFE Preservation Model : key elements explained

Preservation cost of n objects of a particular format for the period 0 to t.

Eg. 200000 objects of the GIF format for a period of 10 years.

Preservation = t * TEW + (t / ULE + PON) * (CRS + UME + PPA + QAA)

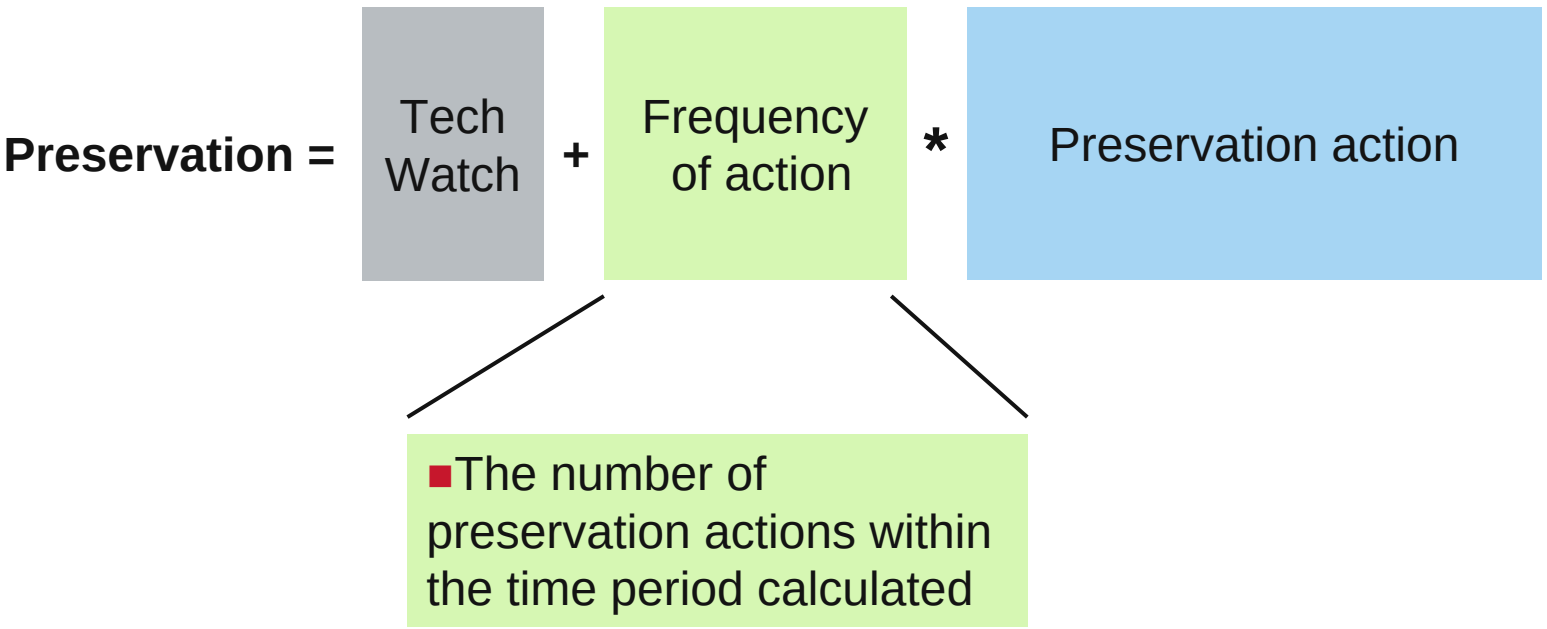


The Generic LIFE Preservation Model : key elements explained

Preservation cost of n objects of a particular format for the period 0 to t.

Eg. 200000 objects of the GIF format for a period of 10 years.

Preservation = t * TEW + (t / ULE + PON) * (CRS + UME + PPA + QAA)

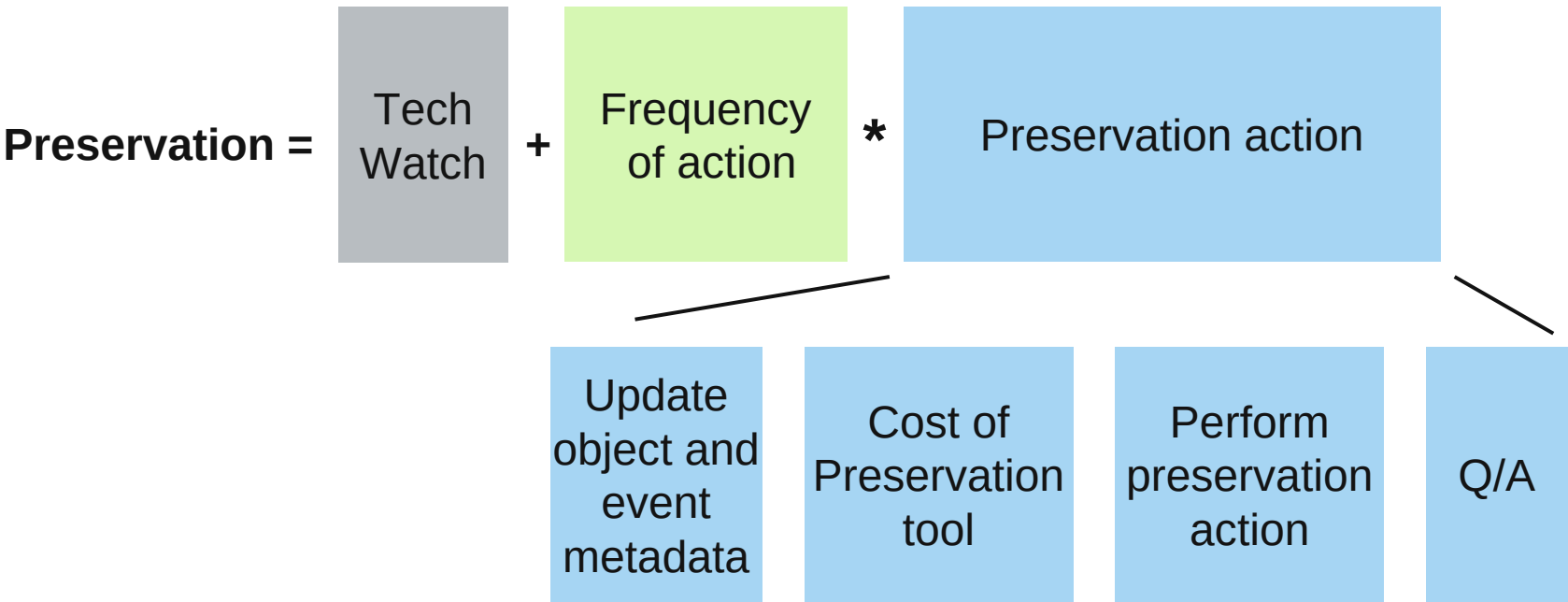


The Generic LIFE Preservation Model : key elements explained

Preservation cost of n objects of a particular format for the period 0 to t.

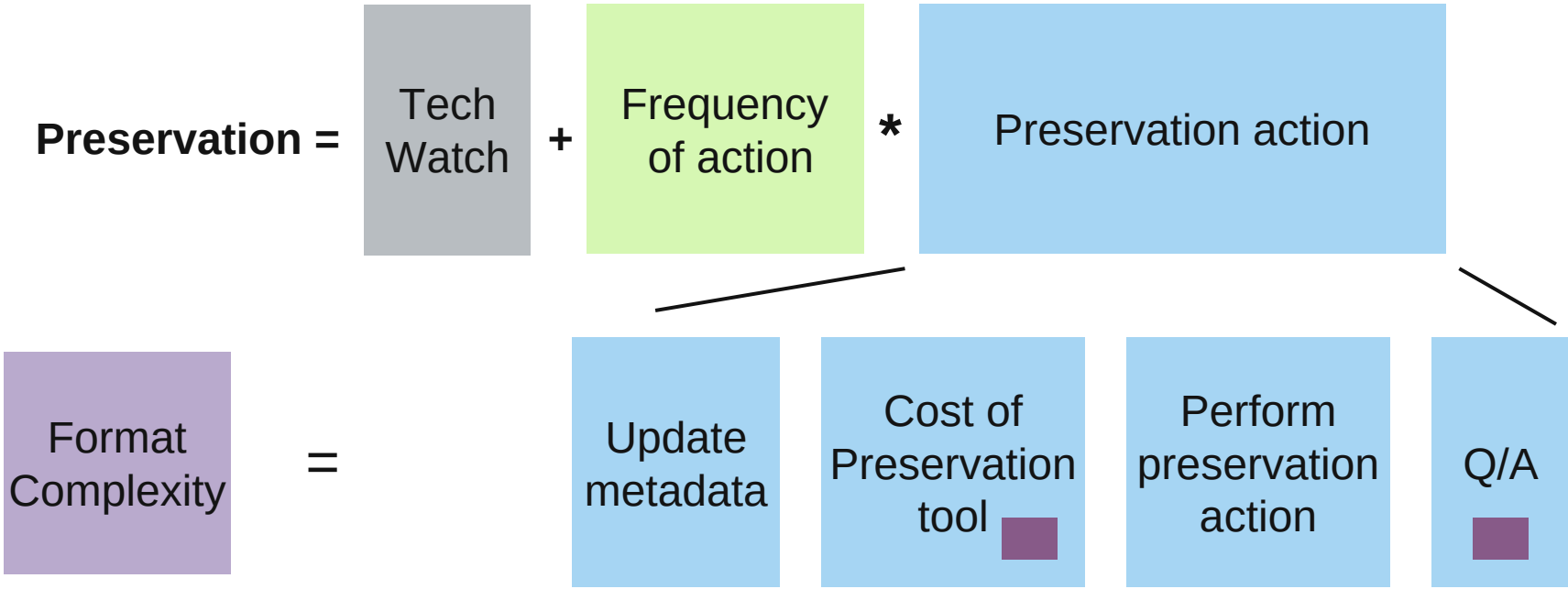
Eg. 200000 objects of the GIF format for a period of 10 years.

Preservation = t * TEW + (t / ULE + PON) * (CRS + UME + PPA + QAA)



LIFE

Complexity of file formats (1st detailed sample of the model)



LIFE

Complexity of file formats (1st detailed sample of the model)

Format Complexity =

- **Size**
- **Complexity**
- **Proprietary**
- **Open**
- **Standardised**

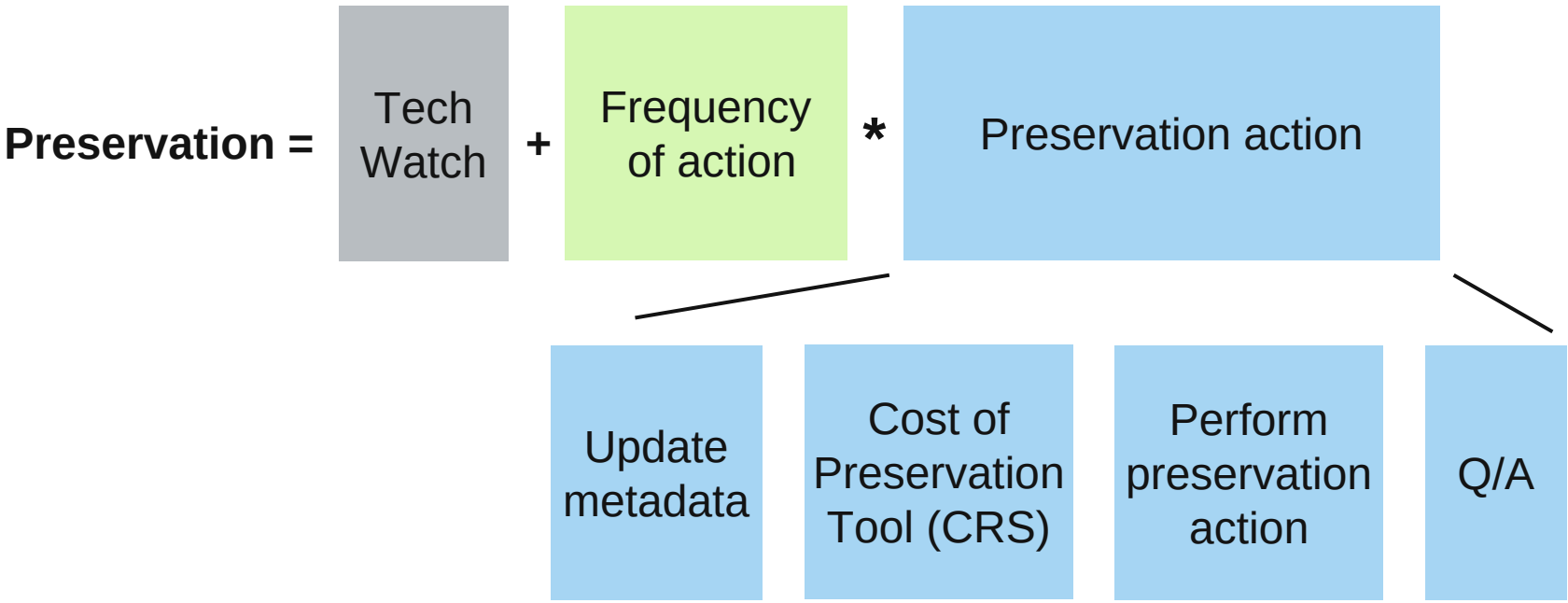
Complexity of file formats (1st detailed sample of the model)

Format Complexity =

Category	Complexity	Examples
Simple	0.1	ASCII, Unicode
Bitmap	0.2	JPEG, GIF
Mark-up	0.3	XML, HTML
Vector	0.4	EMF, Draw
Multimedia	0.6	MPEG3, WAV
Document	0.8	Word, PDF
Complex	1	Oracle database dump

LIFE

Preservation tool cost (2nd detailed sample of the model)

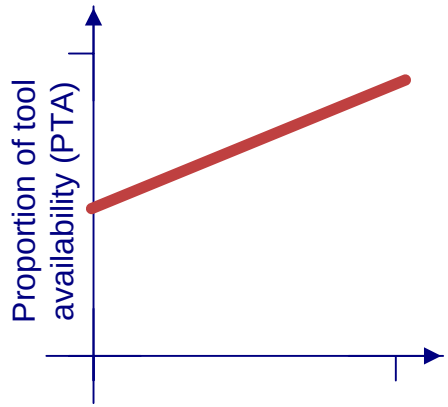


LIFE

Preservation tool cost (2nd detailed sample of the model)

Cost of Preservation Tool (CRS) = Cost of developing a new tool + Cost of acquiring an existing tool

Proportion of tool Availability (PTA)

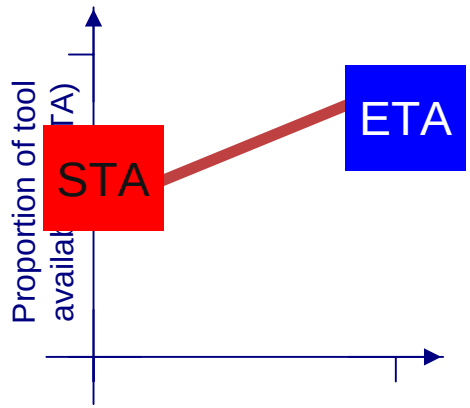


LIFE

Preservation tool cost (2nd detailed sample of the model)

Cost of Preservation Tool (CRS) = Cost of developing a new tool + Cost of acquiring an existing tool

Proportion of tool Availability (PTA)



ETA = 0.9

STA = 0.5

LIFE

Preservation tool cost (2nd detailed sample of the model)

Cost of Preservation Tool (CRS) = Cost of developing a new tool | Cost of acquiring an existing tool

Proportion of tool Availability (PTA) = STA (1-t/20) + ETA (t/20)

LIFE

Preservation tool cost (2nd detailed sample of the model)

Cost of Preservation Tool (CRS) = Cost of developing a new tool | Cost of acquiring an existing tool

Proportion of tool Availability (PTA) = Average proportion across the time period

LIFE

Preservation tool cost (2nd detailed sample of the model)

Cost of Preservation Tool (CRS) = (1- PTA) Cost of developing a new tool + PTA Cost of acquiring an existing tool

Tool Development Cost (TDC) = Estimated as 24 programmer months @ 30k annual salary (£60000)

LIFE

Preservation tool cost (2nd detailed sample of the model)

Cost of Preservation Tool (CRS) = (1- PTA) Tool Development Cost (TDC) Format Complexity + PTA Cost of acquiring an existing tool

Cost of Available tool = Estimated as £1500

LIFE

Preservation tool cost (2nd detailed sample of the model)

Cost of Preservation Tool (CRS) = (1- PTA) Tool Development Cost (TDC) Format Complexity + PTA Cost of Available tool

Preservation tool cost (2nd detailed sample of the model)

Cost of Preservation Tool (CRS) = (1- PTA) Tool Development Cost (TDC) Format Complexity + PTA Cost of Available tool

Preservation = t * TEW + (t / ULE + PON) * (CRS + UME + PPA + QAA)

Case studies

- **LIFE1: Costed and published lifecycles:**
  - Voluntary Deposit Material at the British Library
  - Web Archiving Material at the British Library
  - eJournals at UCL
  - All costs published online at www.life.ac.uk
- **LIFE2: New case studies, just underway:**
  - SHERPA LEAP - 3 institutional repositories
  - SHERPA DP – preservation services
  - Medical Research Council – primary data
  - Burney Collection (BL) – newspapers and digitisation

LIFE2 deliverables

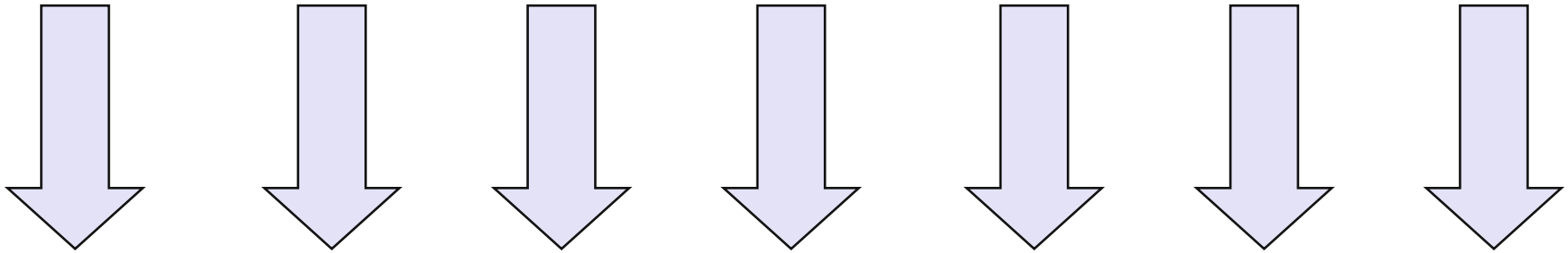
- Report on independent evaluation by economics expert (end 2007)
- Revision of the LIFE Model
  - Version 1.1 (October 2007)
  - Version 2 (summer 2008)
- A detailed and prescriptive methodology for costing digital lifecycles (summer 2008)
- Version 2 of the Generic LIFE Preservation Model (summer 2008)
- Final report, describing 5 new case studies with detailed lifecycle costings (summer 2008)
- End of project conference (summer 2008)

In conclusion...

- More to do, but LIFE techniques are already showing potential in enabling:
  - Improved assessment of the financial commitment an organization is making when acquiring or creating new digital materials.
  - More effective planning for future preservation activities.
  - Comparison of digital lifecycles across an organisation or between different types of organisation.
  - Evaluation and optimisation of existing digital lifecycles.
  - Generation of guidance to funding bodies, such as JISC, to address the aspects of the digital lifecycle which would most benefit from an investment in tool development and automation.

Looking ahead: “LIFE3”

Predictive models for each stage of the lifecycle



Thank you! Questions....?

www.life.ac.uk

LIFE: Costing the Digital Preservation Lifecycle

Paul Wheatley

Digital Preservation Manager

The British Library

Acknowledgements:

Helen Shenton (BL), Rory Mcleod (BL), Paul Ayris (UCL) Adam Farquhar (BL).

Thanks also to Ulla Bøgvad Kejser (Royal Library, Denmark)