Cost models for digitisation and storage of audiovisual archives

(also known as the part of the PrestoSpace experience)

26 July 2005

Matthew Addis (IT Innovation)
Ant Miller (BBC)
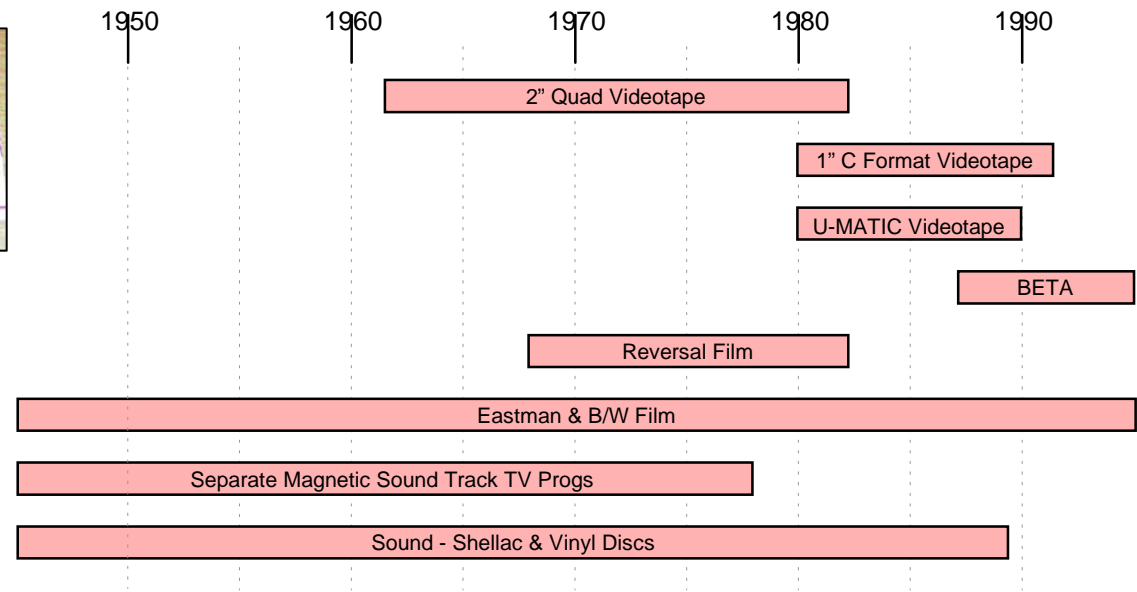
AP1 Alain PERRIER, 13/01/2005

Overview

- Challenges for large audiovisual archives

- The need for planning and cost models

- Mapping using a statistical approach

- Difficult media and long term predictions

- Cost models and projections

- Digital archives

- Summary
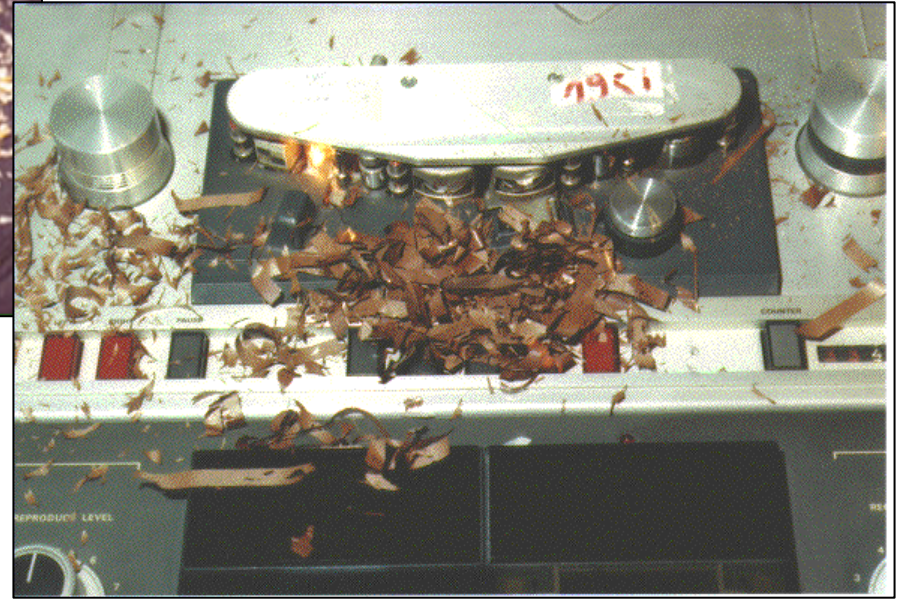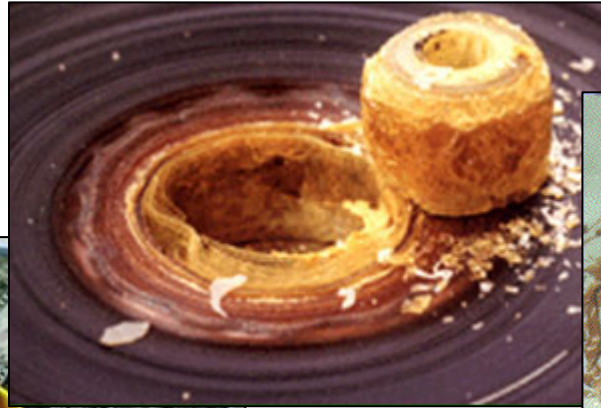
Large Audiovisual Archives

- PrestoSpace estimate: 6M hrs across 20 major European archives
- UNESCO estimate: 200M hrs of film and video in total

PrestoSpace

	1950	1960	1970	1980	1990

2" Quad Videotape

1" C Format Videotape

U-MATIC Videotape

BETA

Reversal Film

Eastman & B/W Film

Separate Magnetic Sound Track TV Progs

Sound - Shellac & Vinyl Discs

VHS

DVCAM

DVCPro

MiniDV

Digital

DigiBeta

DVD

CD-R CD Recordable 650 MB 74 MIN

CD-ROM

Mini CD/DVD

At least 2/3 of the material cannot be easily used

- Approx 1/3 of material has deterioration
- Approx 1/4 of material cannot be released as it is too easily damaged

The need for a cunning plan

- 10 to 20 years is not uncommon for a preservation project
- PrestoSpace Survey
  - 250,000 items per year at a cost of 30M Euro
  - This is still only 1.5% of total holdings each year!
  - Not enough money, capacity, time
- Loss due to decay and obsolescence is inevitable
- Best case, 40% of tape based content will be lost by 2045
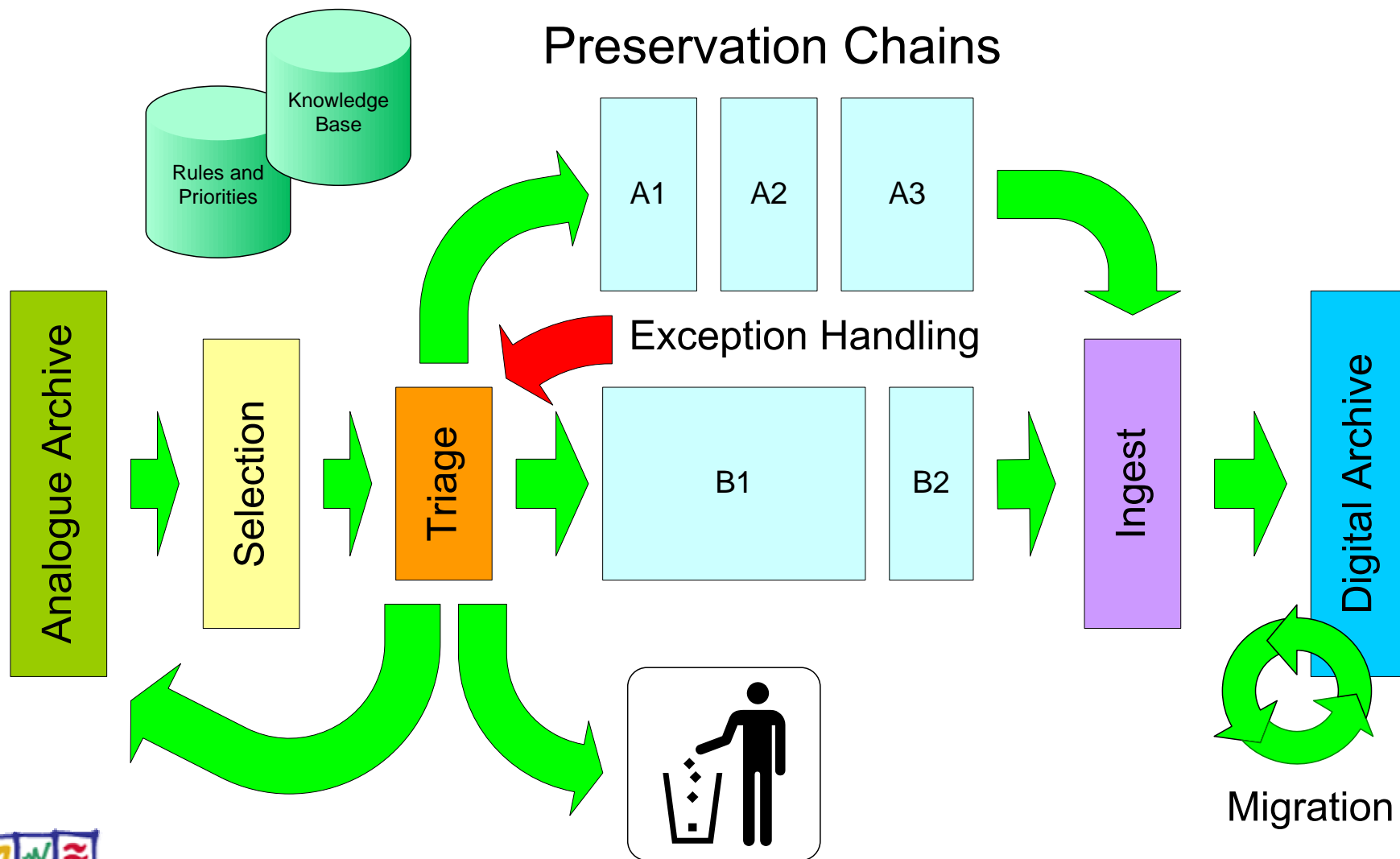- Worst case, 70% of tape based content will be lost by 2025

Objective

- Help archive managers to plan the digitisation and storage of large audiovisual collections

  - How much will it cost?
  - How long will it take?
  - How much will be lost?
  - What should be done first?
  - What can wait until later?
  - What workflows should be used?

Approach

- **Work out what you have**
  - Technical map (carriers, formats, conditions)
  - Content map (genres, value)
  - Use a statistical approach
- **Work out your priorities for preservation**
  - Value of information assets
  - Model what will happen as a function of time
  - Optimise preservation in terms of cost/quality/volume/loss
- **Use an efficient workflow**
  - Triage, sorting, selection
  - Preservation chains and exception handling
  - Knowledge bases to improve decision making
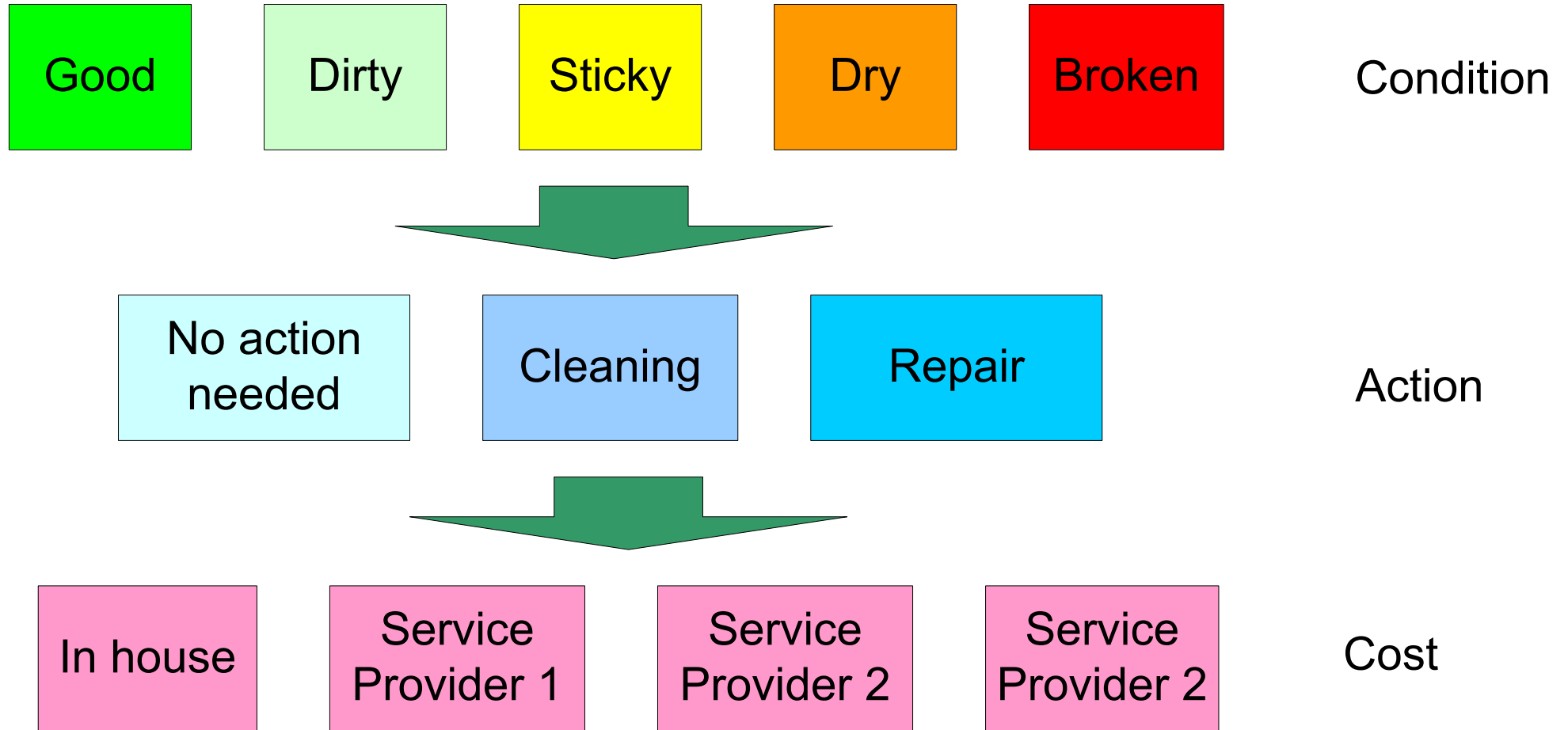 - Migration within the digital archive
- **Make year on year preservation plan**

Workflow

Preservation Chains

Knowledge Base

Rules and Priorities

A1 A2 A3

Exception Handling

Analogue Archive

Selection

Triage

B1 B2

Ingest

Digital Archive

Migration

Workflow

- **Triage based assessment of batches and items**
  - Condition
  - Cataloguing

- **Identify simple tests and measurements**
  - Simple chemical markers, e.g. A-D strips
  - Visual inspection, e.g. media and containers (cassettes, reels)
  - Mechanical tests, e.g. rewinding, clogging, playback

- **Create a knowledge base**
  - Serial numbers $\rightarrow$ condition prediction $\rightarrow$ cost prediction

- **Reject unplayable items**
  - Don't waste time attempting transfer

- **Allocation of items to preservation chains**
  - Minimise 'exceptions' in expensive stages
  - Avoid damage to machines
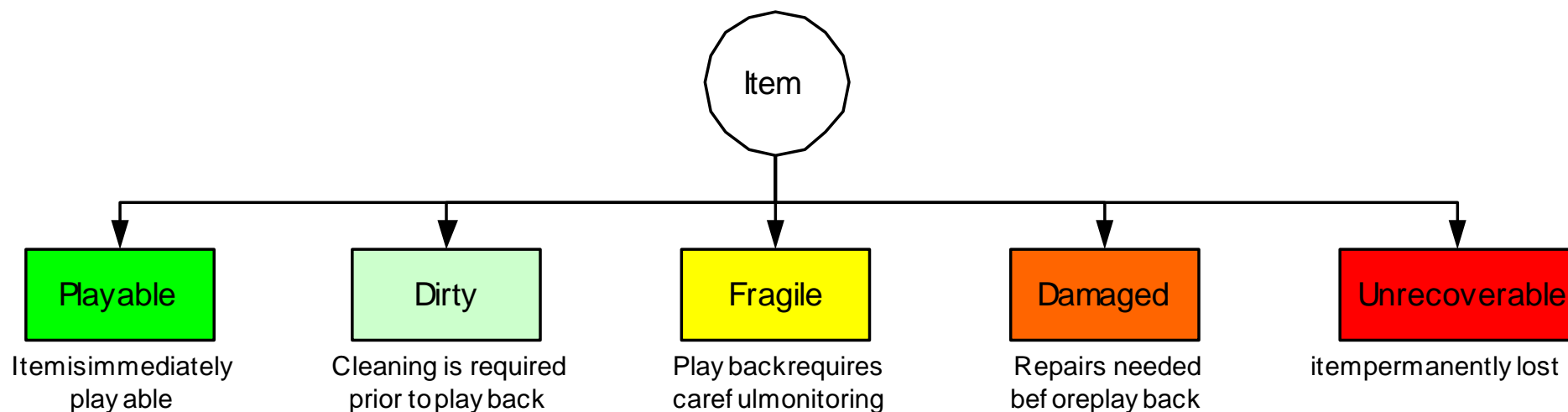
Mapping the archive

- **Impractical to map the entire archive**
  - Media condition and content typically not known until items are taken off the shelf
  - Takes too long, costs too much

- **Take a sample and use statistics**
  - Direct investigations and pilot studies
  - Indirect picture from user experiences
  - Estimate the overall status of the archive for planning purposes
  - But can't tell you in advance what to do for each item

Media condition

- **Chemical state**
  - Vinegar syndrome, binder hydrolysis, lubricants and additives
  - Splices, leaders

- **Physical condition**
  - Broken sprockets, shrinkage, scratches
  - Stretching, creases, wear and tear
  - Damage to cassettes and reels
  - Mould, dirt

- **Multiple factors can be present**
  - Chemical decay + wear and tear + accidental damage
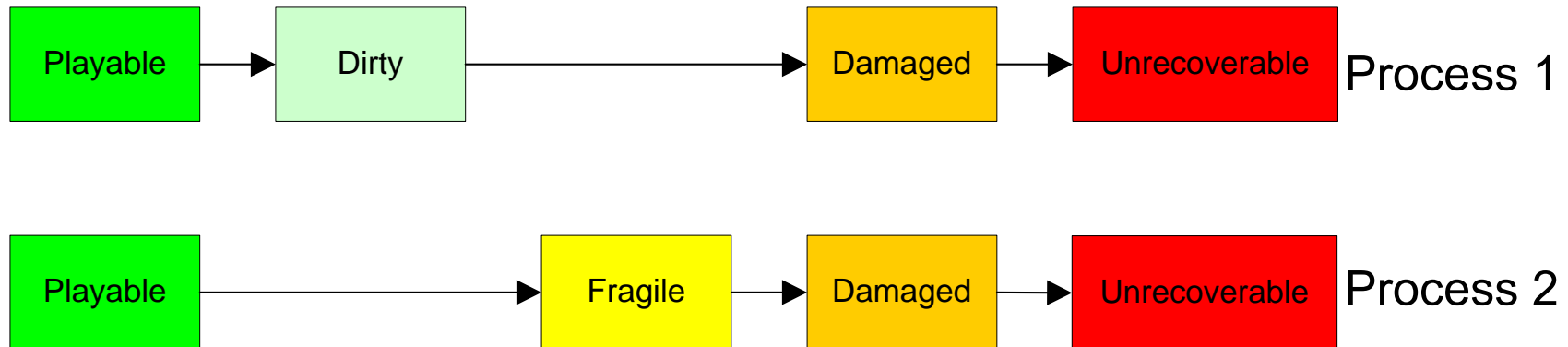
Mapping from condition to cost

| Good | Dirty | Sticky | Dry | Broken | Condition |

| No action needed | Cleaning | Repair | Action |

| In house | Service Provider 1 | Service Provider 2 | Service Provider 2 | Cost |

Modelling media condition

```
                              ┌──────────┐
                              │   Item   │
                              └────┬─────┘
        ┌───────────┬─────────────┼─────────────┬───────────────┐
        ▼           ▼             ▼             ▼               ▼
   ┌─────────┐  ┌─────────┐  ┌─────────┐  ┌─────────┐  ┌──────────────┐
   │Playable │  │  Dirty  │  │ Fragile │  │ Damaged │  │Unrecoverable │
   └─────────┘  └─────────┘  └─────────┘  └─────────┘  └──────────────┘
```

Playable	Dirty	Fragile	Damaged	Unrecoverable
Item is immediately play able	Cleaning is required prior to play back	Play back requires careful monitoring	Repairs needed before play back	item permanently lost

Carrier Name	Playable % of carrier	Dirty % of carrier	Fragile % of carrier	Damaged % of carrier	Unrecoverable % of carrier	% of collection
2" Quad	15%	45%	30%	7%	3%	15%
1" C Format	32%	45%	13%	6%	3%	3%
¾ " UMatic	27%	13%	54%	1%	5%	82%

Task list			Services		
Carrier		**Items**	**Provider**	**Cost now**	**Capacity**
Carrier	*Condition*	*Unit*	*Name*	*Euros/Unit*	*Units/Year*
2" Quad	Playable	4594	Company A	200	1500
			Company C	220	1600
			Company E	250	2500
	Dirty	13782	Company A	240	1200
			Company B	230	1100
			Company C	250	2000
	Fragile	9188	Company B	300	800
			Company C	320	900
			Company D	290	600
	Damaged	2297	Company F	380	400
			Company G	420	600
			Company H	460	900
1" C format	Playable	21165	Company B	140	1600
			Company D	140	1700
			Company F	150	2300
	Dirty	29630	Company A	160	1200
			Company B	175	1400
			Company C	180	1800
	Fragile	8466	Company A	210	1600
			Company B	200	1500
			Company D	230	3000
	Damaged	4233	Company F	260	800
			Company G	280	900
			Company H	290	1200
3/4" Umatic	Playable	14411	Company E	50	3000
			Company G	60	3400
			Company H	55	3200
	Dirty	7206	Company E	60	2000
			Company G	70	1700
			Company H	65	2100
	Fragile	28823	Company C	70	1200
			Company F	75	1300
			Company H	80	1500
	Damaged	288	Company A	100	1000
			Company B	95	900
			Company D	105	1100

Modelling degradation

Playable → Dirty → Damaged → Unrecoverable Process 1

Playable → Fragile → Damaged → Unrecoverable Process 2

Condition	Future Condition				
	Playable	**Dirty**	**Fragile**	**Damaged**	**Unrecoverable**
Current Condition	*% of condition*	*% of condition*	*% of condition*	*% of condition*	*% of condition*
Playable	90%	10%	0%	0%	0%
Dirty		80%	15%	5%	0%
Fragile			70%	20%	10%
Damaged				60%	40%
Unrecoverable					100%

Year 1
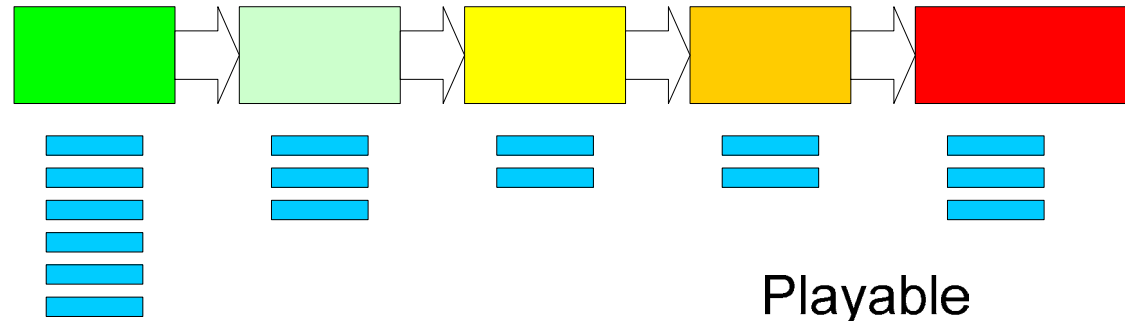
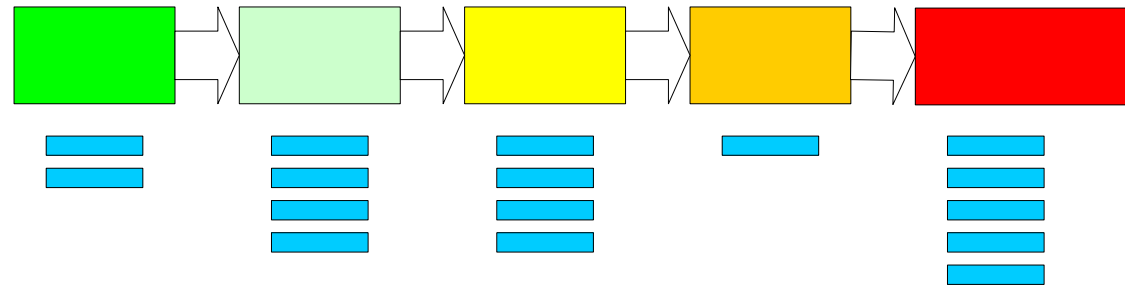Playable   Dirty

Year 2
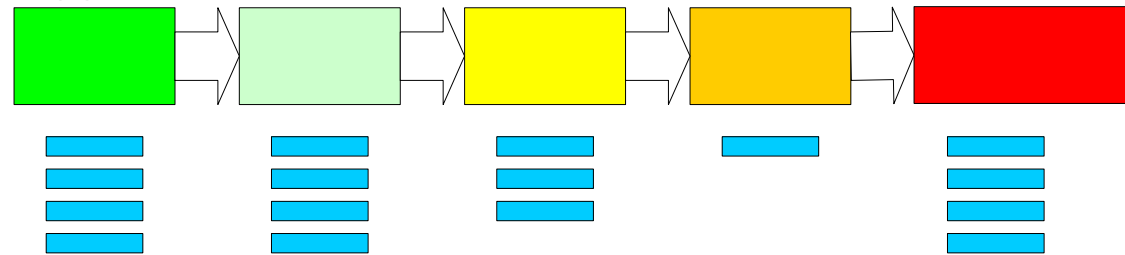
Year 3

Playable   Dirty

Histogramme Vinaigre MARS

Acidité 0   Acidité 1   Acidi

200h

150h

100h

50h

10h

53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87
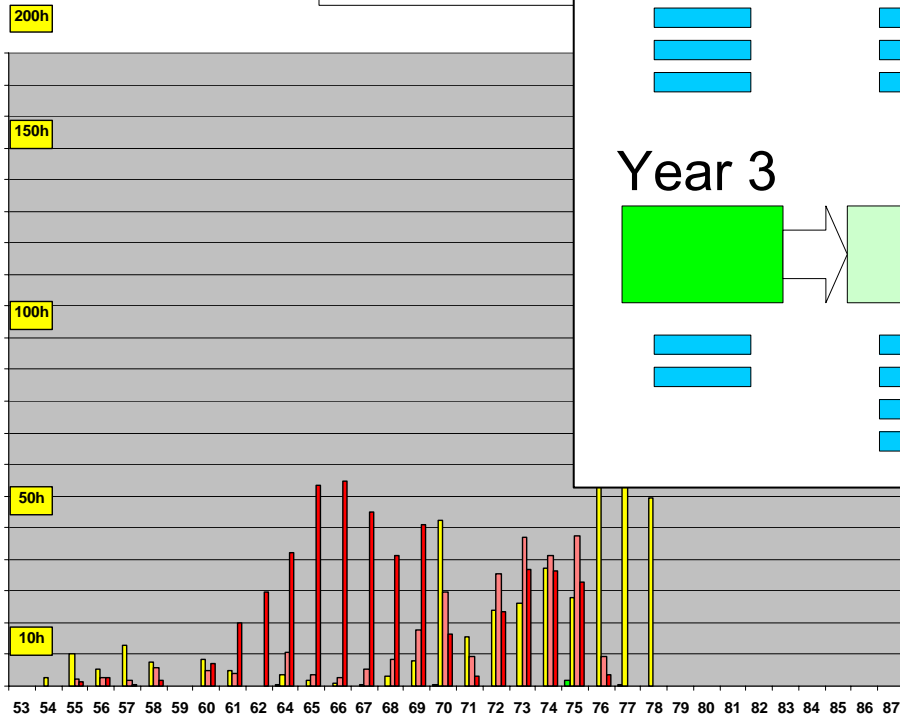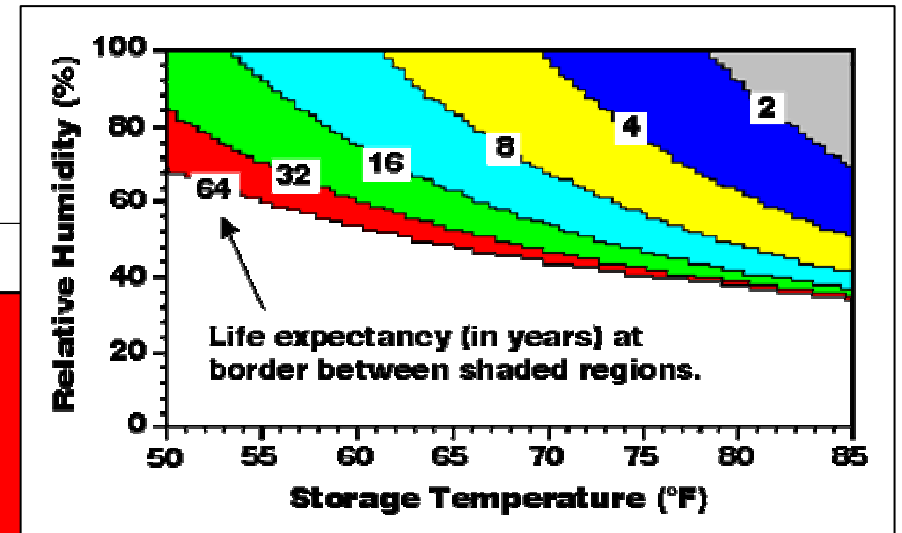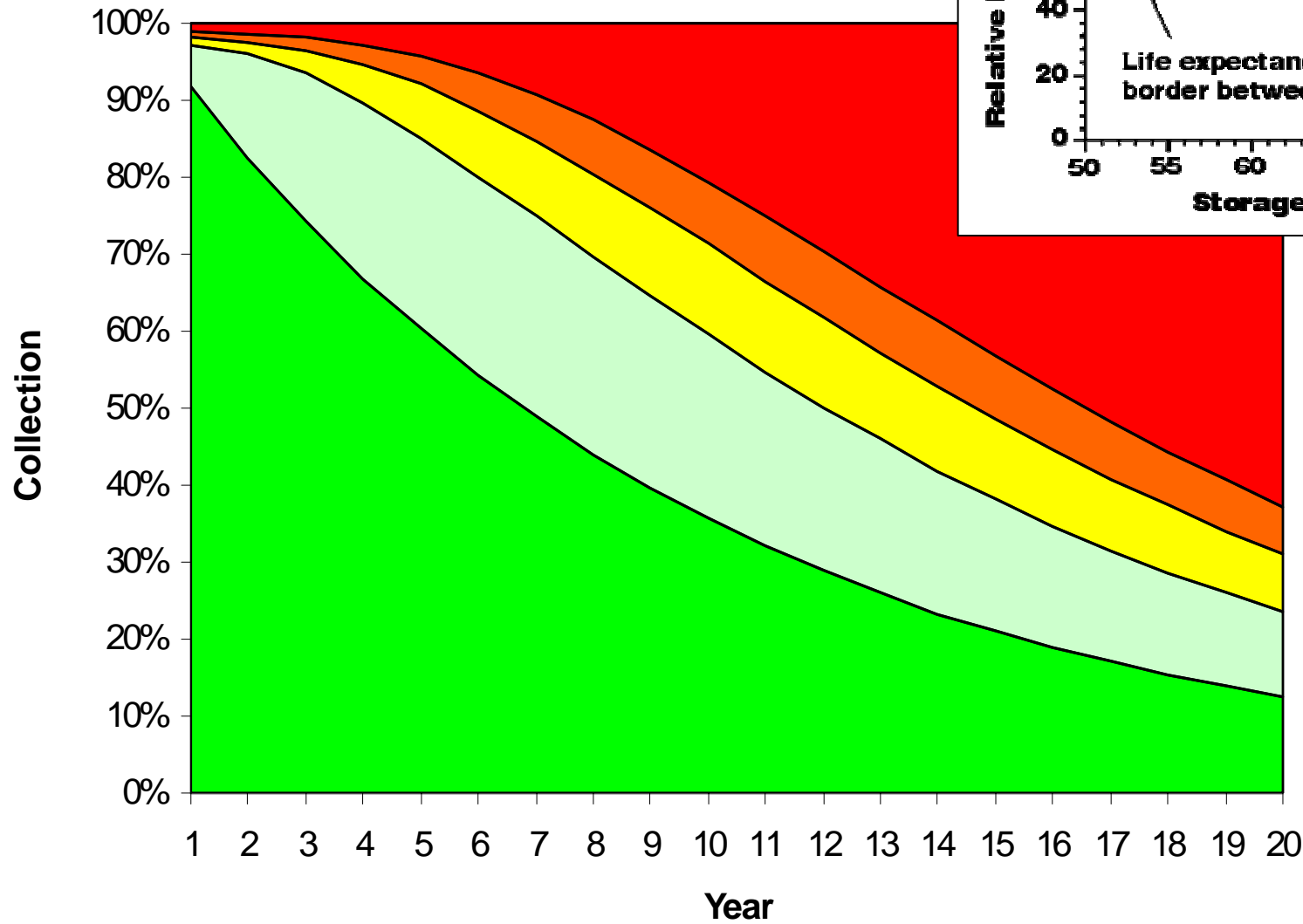
Content mapping

Genre	Collection	Loaned	Popularity	Ranking	Star items	Low Value
Name	Items	Items	*% of collection used each year*		% of genre	% of genre
News	100000	30000	**30%**	2	10	8
Sport	50000	15000	**30%**	3	3	15
Drama	30000	10000	**33%**	1	20	5
Natural History	20000	5000	**25%**	4	3	4
Entertainment	10000	100	**1%**	5	10	8

| Genre	Collection	Loaned	Popularity	Ranking	Preserve	Discard
Name	Items	Items	*% of collection used each year*		% of genre	% of genre
News	100000	30000	**30%**	2	92	8
Sport	50000	15000	**30%**	3	85	15
Drama	30000	10000	**33%**	1	95	5
Natural History	20000	5000	**25%**	4	3	97
Entertainment	10000	100	**1%**	5	10	90

Prioritisation

- Determines order in which items will be processed
- Provides rules for sorting and selection
- Various strategies
  - Most 'valuable' first
  - Worst condition first
  - Obsolete carriers first
  - *Best condition first*
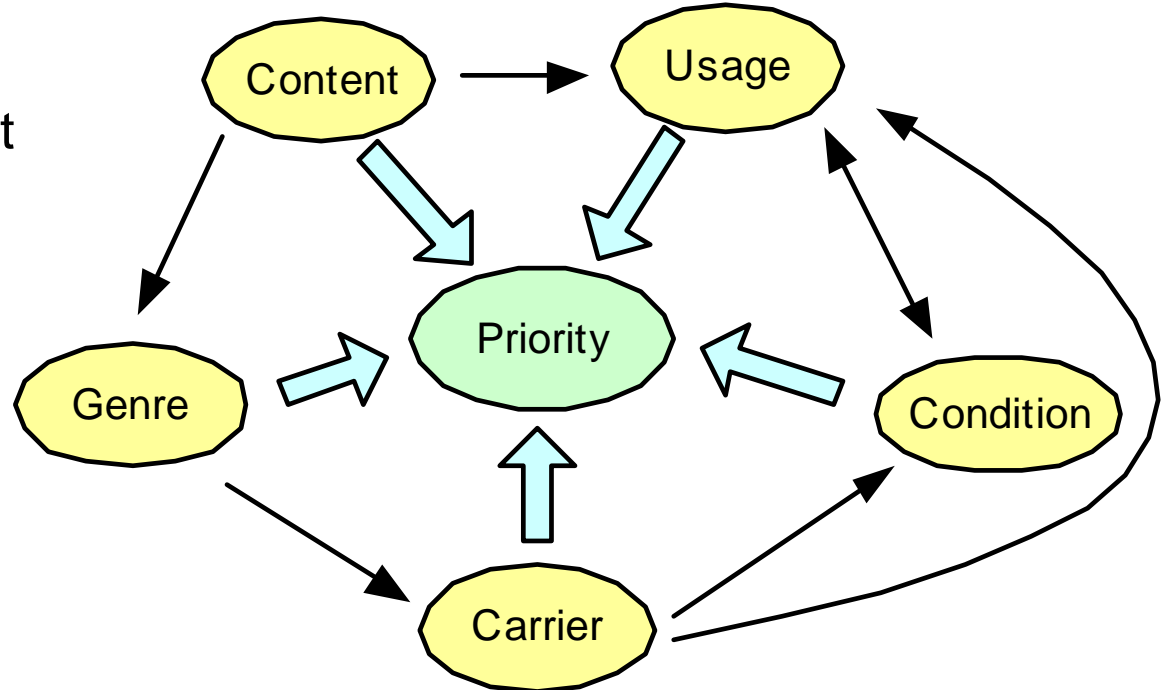
Investigating the options

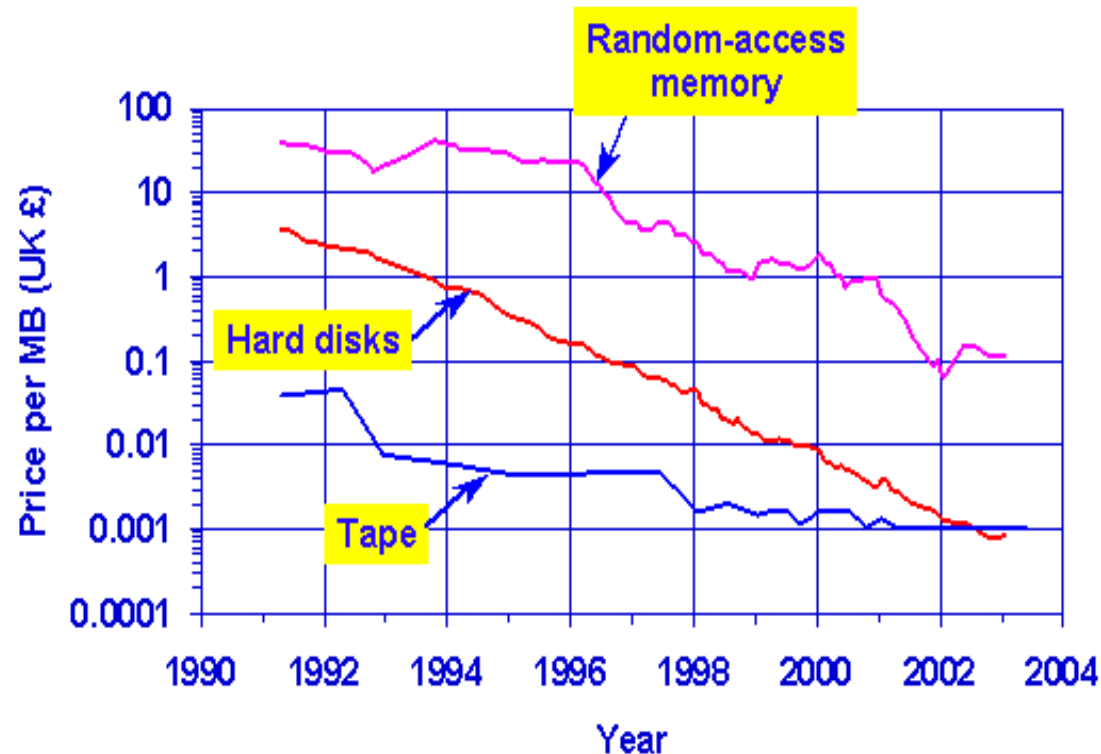| Service Carrier Name	Condition Type	Provider Name	Transfer plan per year										Remains
			0 Items	1 Items	2 Items	3 Items	4 Items	5 Items	6 Items	7 Items	8 Items	9 Items	Items
1" C format Lost 6971	Playable	Company A Company C Company E	1500	1440									0
	Dirty	Company A Company B Company C	1100	1100	1100	1100	523						0
	Fragile	Company B Company C Company D	600	600	600	600	600	433					0
	Repairable	Company F Company G Company H	400	400	400	400	400	400	400	400	400	275	0
¾" umatic Lost 2891	Playable	Company B Company D Company F	1700	1700	1700	1700	1083						0
	Dirty	Company A Company B Company C	1200	1200	1200	1200	1200	1200	371				0
	Fragile	Company A Company B Company D	1500	1500	1500	1500	1500	131					0
	Repairable	Company F Company G Company H	800	800	800	800	800						0
2" Quad Lost 3926	Playable	Company E Company G Company H	3000	3000	1190								0
	Dirty	Company E Company G Company H	2000	2000	286								0
	Fragile	Company C Company F Company H	1200	1200	1200	1200	1200	1200	611				0
	Repairable	Company A Company B Company D	0 900	900	900	900	900	900	410				0
Total cost (Euros)			2258700	2311835	1870536	1842884	1651583	771230	349953	186941	192549	136349	

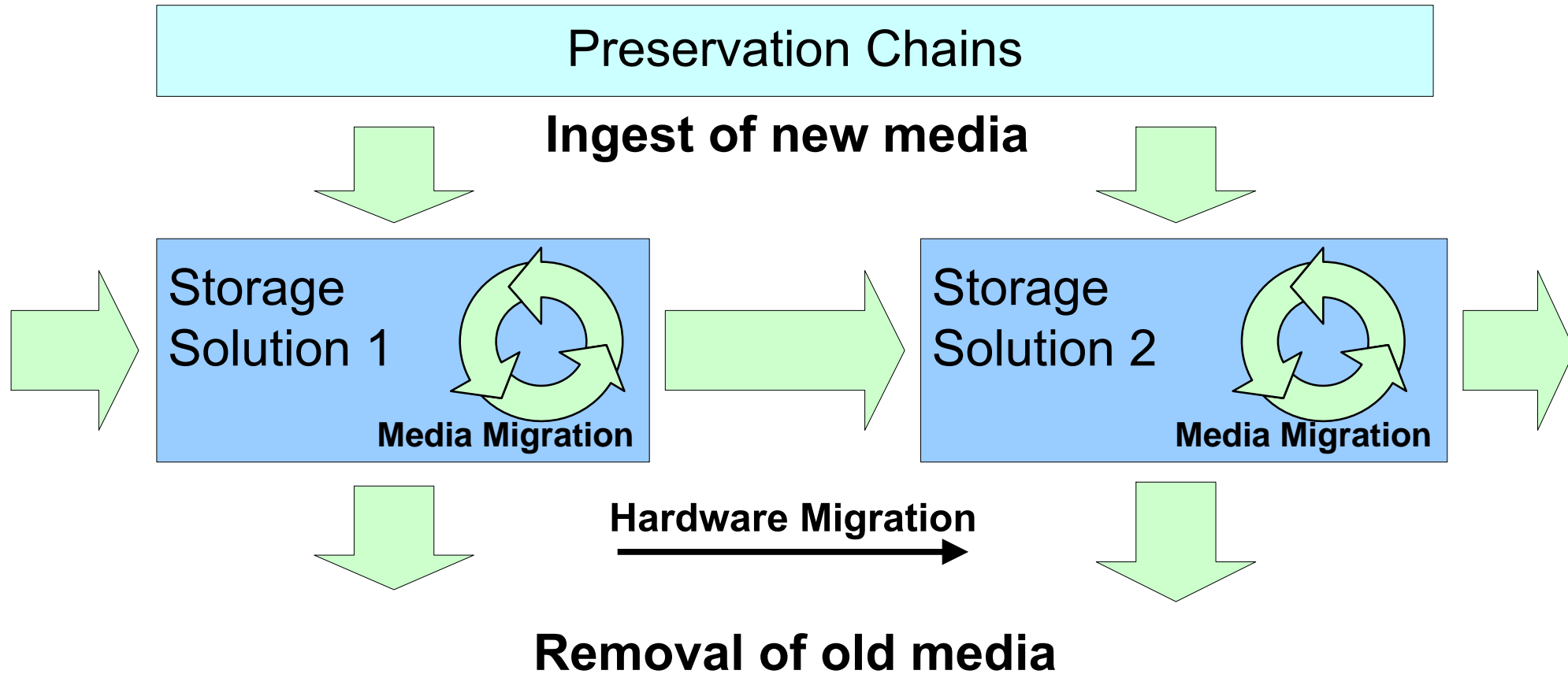# Projections

Digital Archive
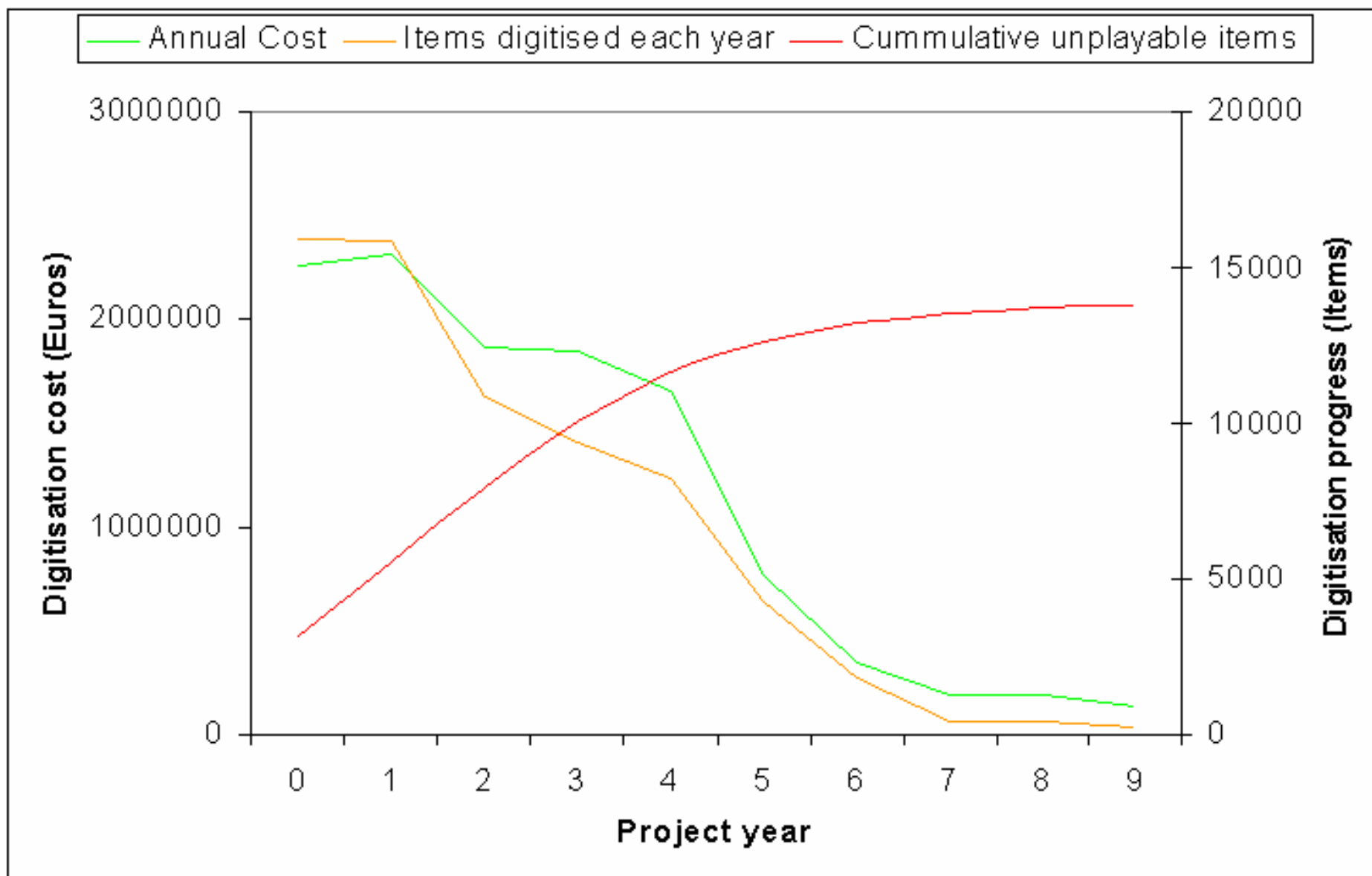
- **Technical obsolescence happens faster**
  - Media discontinued more rapidly
  - Rapid advances in disks, robots, OS, network
  - Different cycles for file formats and media types
  - Change storage systems as often as every 3 to 5 years

- **Moore's law**
  - Rapidly falling storage costs (hardware, space, media)
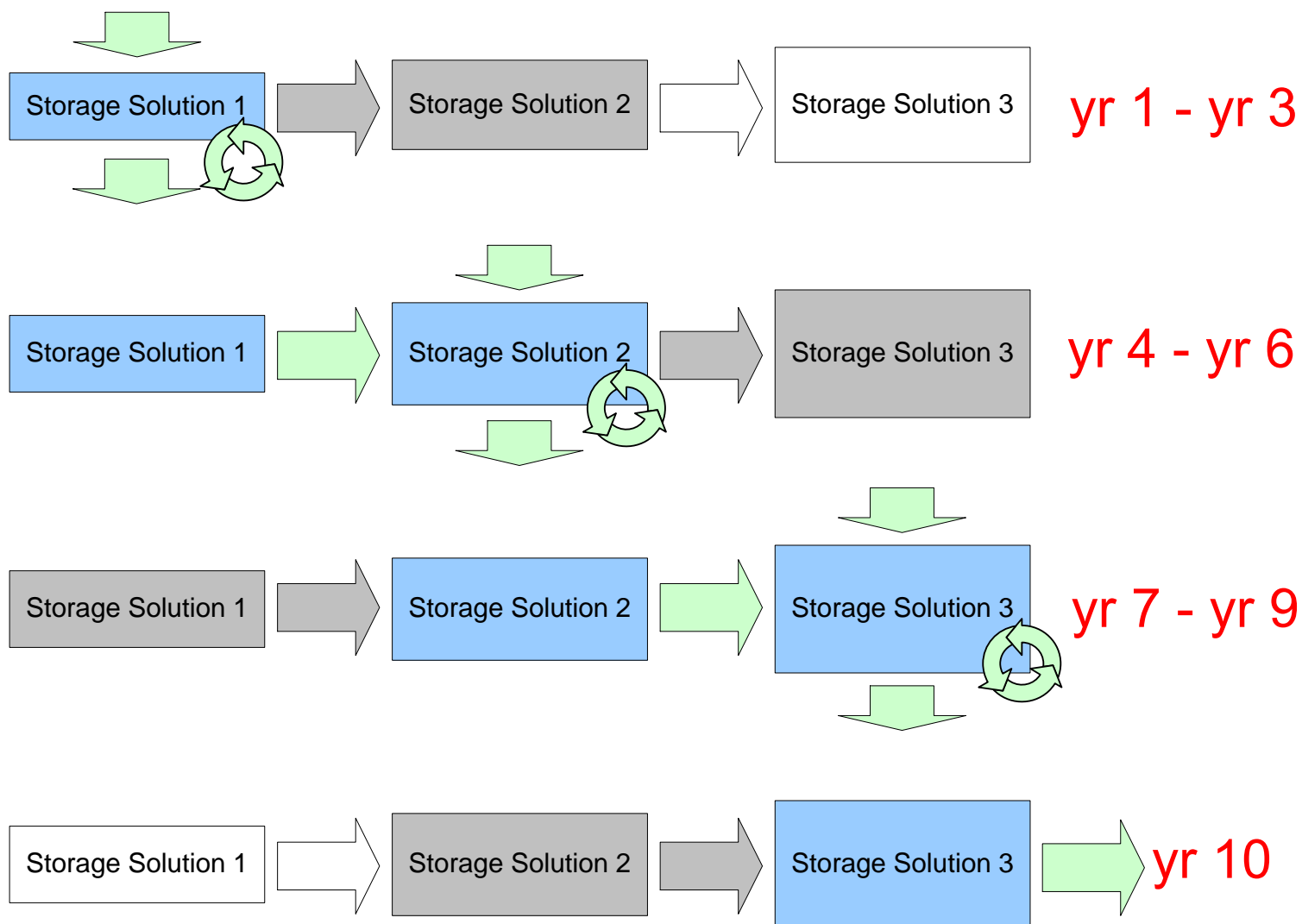 - Faster access, move towards online systems

- **Off the shelf solutions**
  - Not specific to broadcasting

Digital archive model

Migration plan

| Storage Solution 1 | → | Storage Solution 2 | ⇒ | Storage Solution 3 | yr 1 - yr 3 |

| Storage Solution 1 | ⇒ | Storage Solution 2 | → | Storage Solution 3 | yr 4 - yr 6 |

| Storage Solution 1 | → | Storage Solution 2 | ⇒ | Storage Solution 3 | yr 7 - yr 9 |

| Storage Solution 1 | ⇒ | Storage Solution 2 | → | Storage Solution 3 | yr 10 |

Solution	Equipment		
	Capacity	**Longevity**	**Cost**
Name	*Media units*	*Years*	*Euros*
DVD	1000	10	5000
Tape G1	440	3	250000
Tape G2	440	5	250000
Harddisk	40	5	100000

Solution	p(solution'|solution)			
	DVD	**Tape G1**	**Tape G2**	**Harddisk**
Name	*% of solution*	*% of solution*	*% of solution*	*% of solution*
DVD	0%	0%	0%	100%
Tape G1	0%	0%	100%	0%
Tape G2	0%	0%	50%	50%
Harddisk	0%	0%	0%	100%

			Year									
			0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Plan	2" tape	DVD										
		Tape G1	100%	100%	100%							
		Tape G2				100%	100%	100%				
		Harddisk							100%	100%	100%	100%
	1" tape	DVD	100%	100%	100%	100%	100%					
		Tape G1										
		Tape G2										
		Harddisk						100%	100%	100%	100%	100%
	¾" Umatic	DVD										
		Tape G1										
		Tape G2										
		Harddisk	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Media requirements
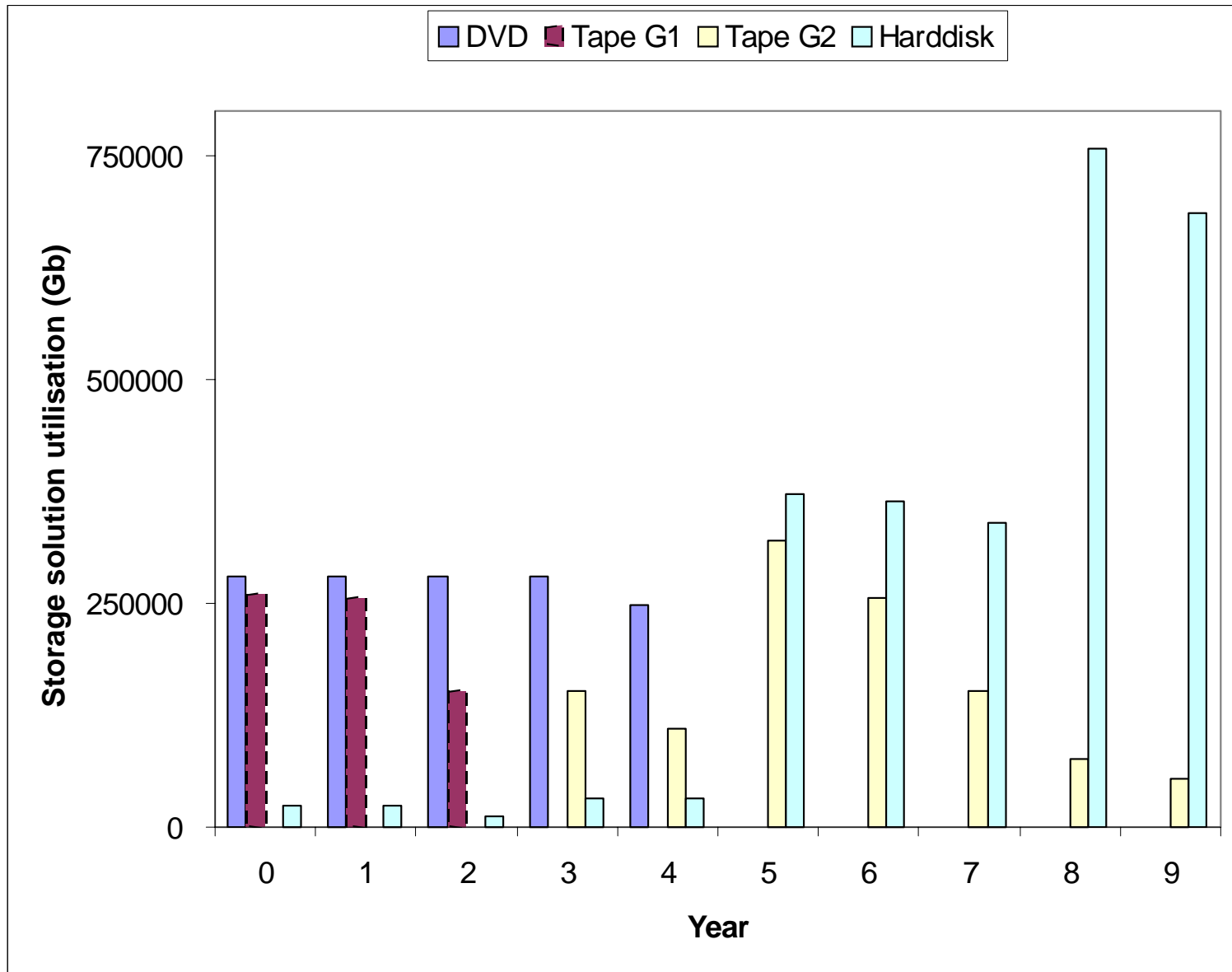
Overall projections
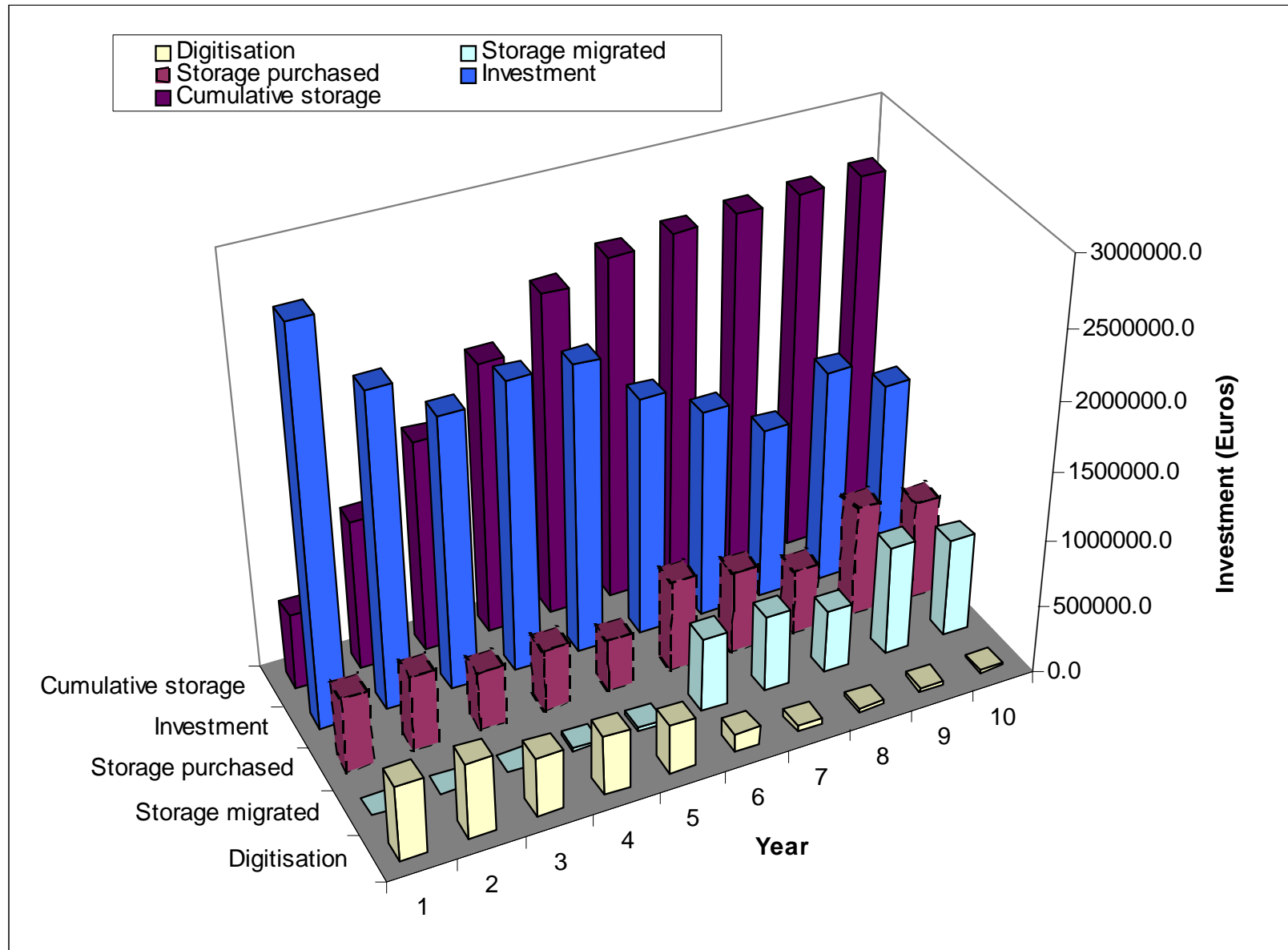
Next steps

- **Calibrate model with real world numbers**
  - Degradation rates, Moore's law, transfer costs, storage costs
  - Check model against existing plans
- **Issue report**
  - September 2005
- **Update report to address needs of small archives**
  - Next two years

Summary

- Broadcast archives face many preservation problems
- Digital archives *could* face many of these problems in the future

- Base cost estimates on statistical models and projections
  - Degradation, obsolescence, inflation
  - Calculate year-on-year costs and losses
  - Investigate trade-offs
  - Can't be specific about individual items $\rightarrow$ needs handling in workflow

- Define digital archive strategy
  - Ongoing migration is more cost effective in the long term
  - Grow the digital archive 'on demand' to reduce upfront costs
  - Watch out if you start putting stuff on the shelf