PREMIS, METS and preservation metadata:

emerging trends and future directions

Eld Zierau
The Royal Library of Denmark

Introduction

My background

- Masters in Computer Science in 1989
- At the Royal Library of Denmark since 2007
  - Strategy and design of preservation systems
  - Creation of preservation policies and strategies
  - Policies of using preservation metadata
- PhD in Digital Preservation in 2011

Currently at the Royal Library

- SIFD – the digital library
  Management, dissemination and preservation
- Packaging and re-packaging for Bit Repository
  WARC, METS, PREMIS
- Framework for OAIS and Distributed Digital Preservation

Contents of this presentation

- **Practices (at The Royal Library)**
  - Strategies and policies
  - Putting it into practice

- **Challenges**
  - Expressing **preservation levels** and **intellectual entities**
  - Preserving preservation metadata
  - Expressing **preservation levels** and **intellectual entities** over time

- **An example on the bit level**
  - Risks mitigated in bit preservation
  - Bit integrity/safety, confidentiality and availability

- **Types of preservation Levels**
  - How to express them – also over time

- **Identification of intellectual entities**
  - How to express them – also over time
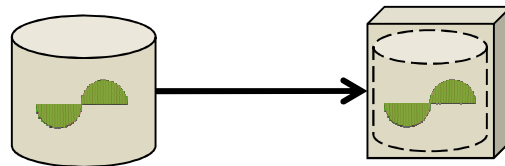
- **Summary**

Preservation Strategies

- **Logical preservation**
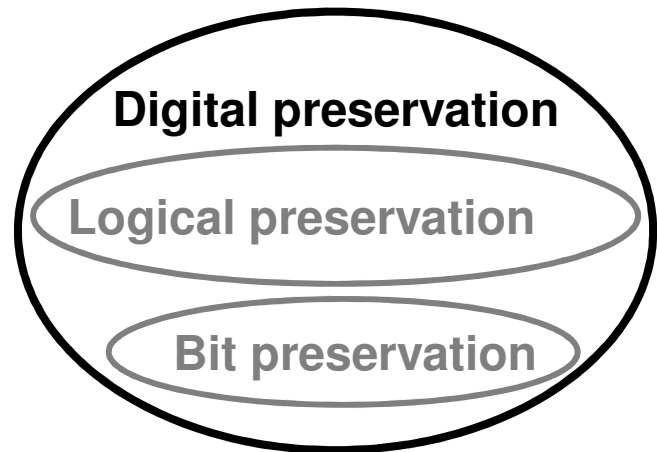
  ◦ Migration
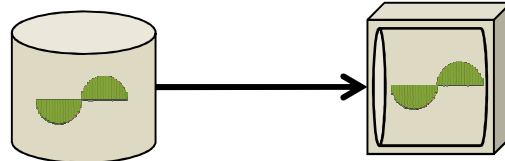
  ◦ Emulation

  ◦ Technology preservation

- **Bit preservation**

  0101100010001000   ...

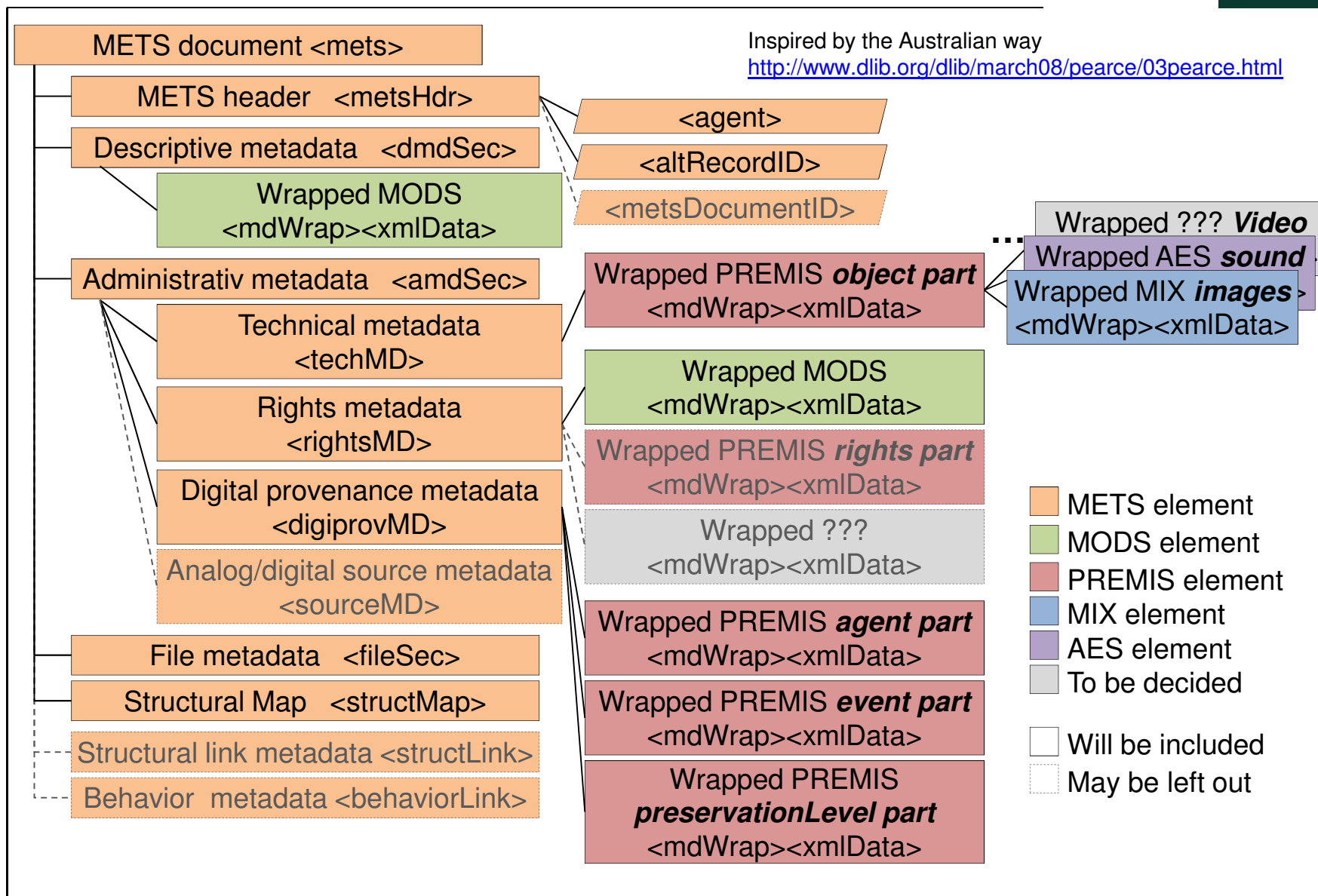Digital preservation

Logical preservation

Bit preservation

Currently at The Royal Library

- Strategy and policies
  - Bit preservation
  - Logical preservation

- Putting it into practice
  - The chosen Metadata Standards
  - The Digital Library infrastructure
  - The Danish Bit Repository Framework

Metadata Standards and use

```
METS document <mets>
    METS header   <metsHdr>
    Descriptive metadata   <dmdSec>
            Wrapped MODS
            <mdWrap><xmlData>
    Administrativ metadata   <amdSec>
            Technical metadata
            <techMD>
            Rights metadata
            <rightsMD>
            Digital provenance metadata
            <digiprovMD>
            Analog/digital source metadata
            <sourceMD>
    File metadata   <fileSec>
    Structural Map   <structMap>
    Structural link metadata <structLink>
    Behavior  metadata <behaviorLink>
```

<agent>

<altRecordID>

<metsDocumentID>

Wrapped PREMIS *object part*
<mdWrap><xmlData>

Wrapped ??? *Video*
Wrapped AES *sound*
Wrapped MIX *images*
<mdWrap><xmlData>

Wrapped MODS
<mdWrap><xmlData>

Wrapped PREMIS *rights part*
<mdWrap><xmlData>

Wrapped ???
<mdWrap><xmlData>

Wrapped PREMIS *agent part*
<mdWrap><xmlData>

Wrapped PREMIS *event part*
<mdWrap><xmlData>

Wrapped PREMIS
preservationLevel part
<mdWrap><xmlData>

Legend:
- METS element
- MODS element
- PREMIS element
- MIX element
- AES element
- To be decided

- Will be included
- May be left out

So what to do

- PREMIS does not do it all (Richard)
  e.g. other standards for technical metadata
- PREMIS – do not re-invent (Richard)
- Sustainability, Community (Huw Jones, Dave)
- Rights "packaging" (All)
  as it suits your organisation
- Material – representations (Angela, Robert)
  AIP be careful – different levels, lot of information
- Events when useful on different levels (audit trail)
- All sorts of agents, challenges with description

- METS why – and on what? (Steffen)
- Large METS (Steffen, Huw Jones)
- Tree-structure – not net (as e.g. web or emails)

Digital Library infrastructure
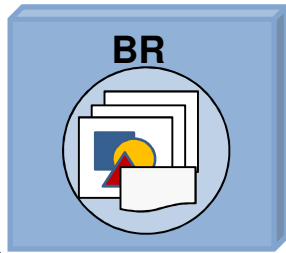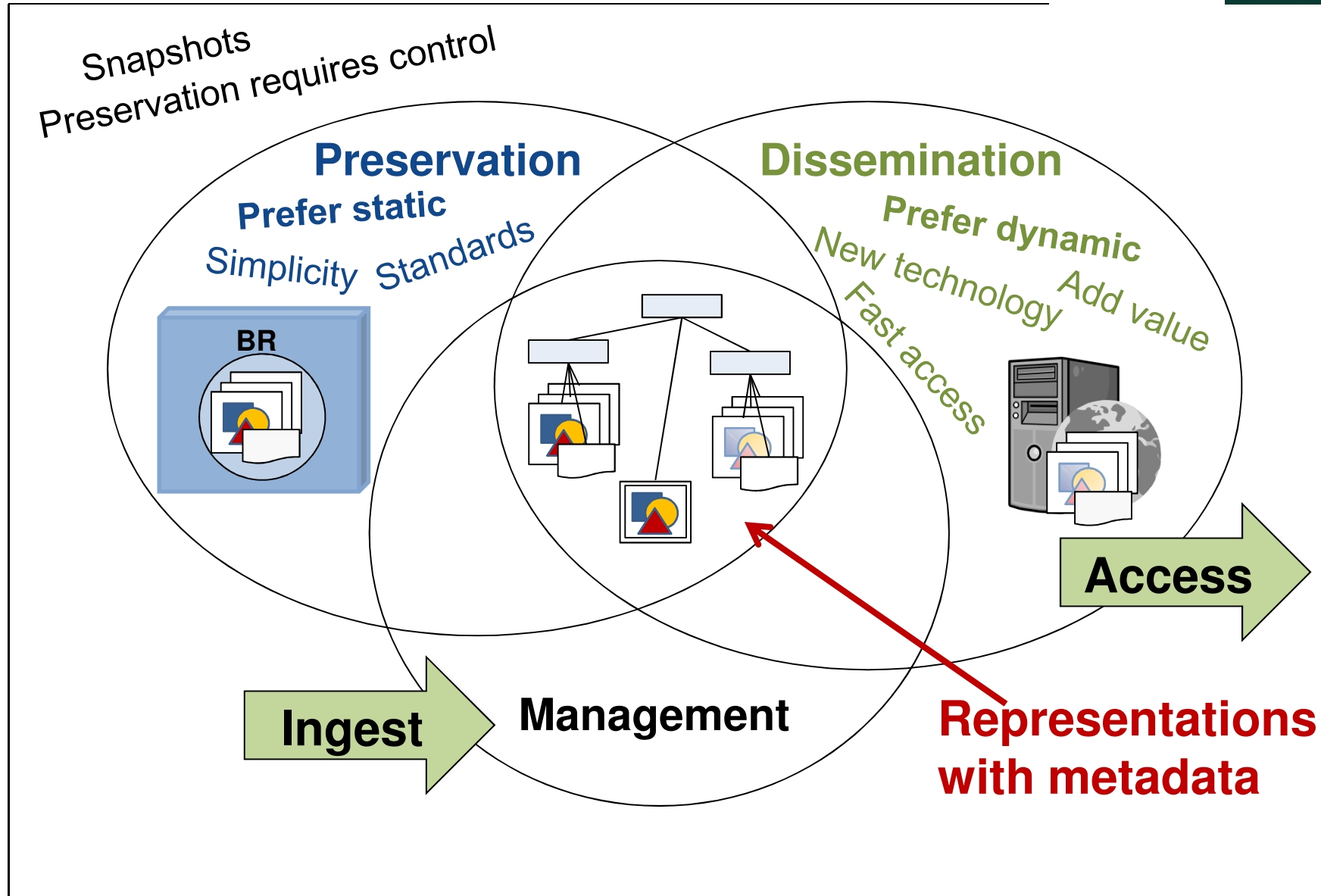
Challenges with metadata

- Preserving preservation metadata (if …)

  Angela write ones ..
  Robert - PREMIS storage ..

- Expressing preservation levels
- Expressing preservation levels over time

  We need to look more closely on bit preservation to define levels and levels over time

  Robert - how many copies .. – more to it

- Expressing Intellectual Entity (identifiers)
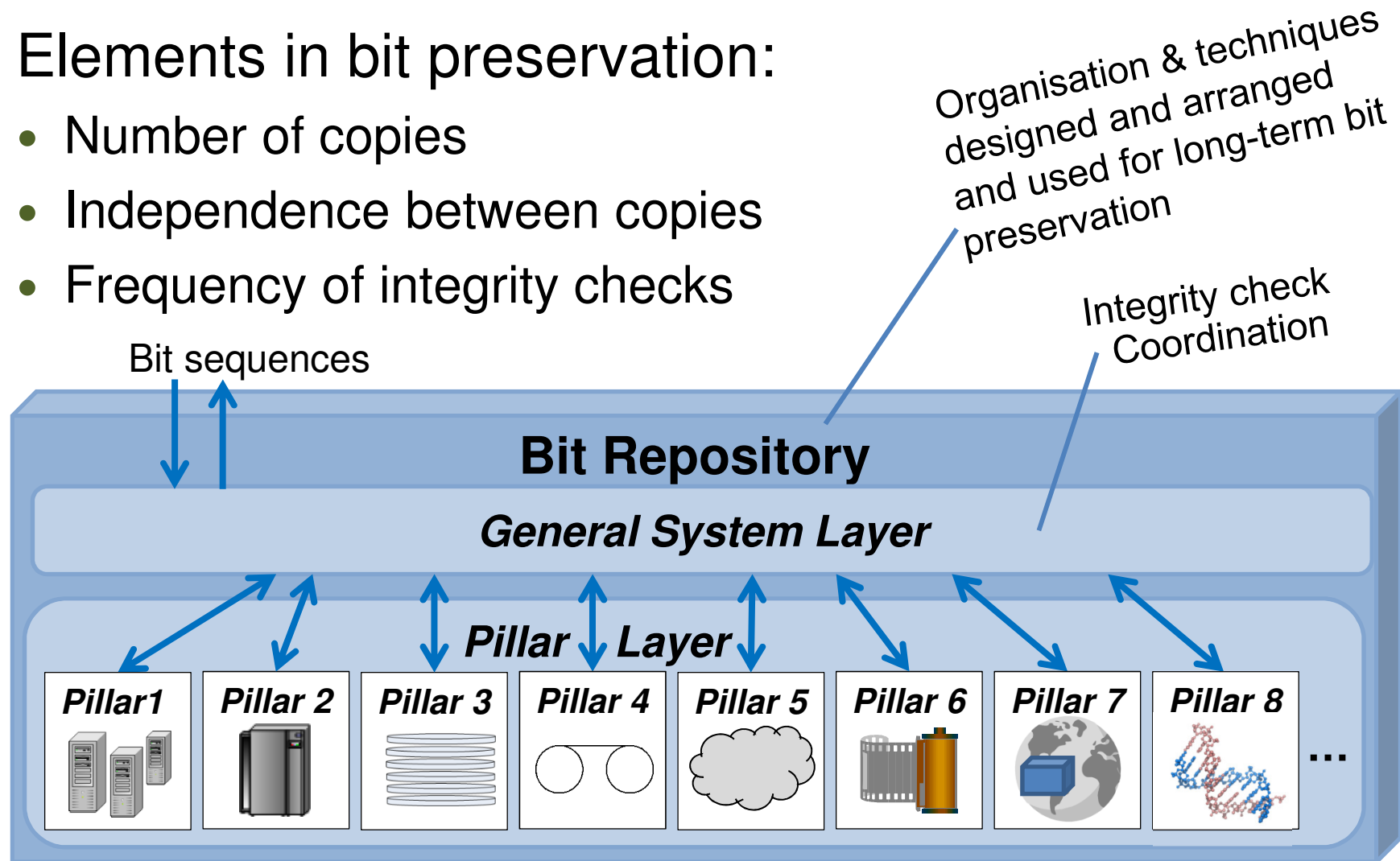- Expressing Intellectual Entity (identifiers) over time

A General View of a Bit Repository

Elements in bit preservation:

- Number of copies
- Independence between copies
- Frequency of integrity checks

Bit sequences

Organisation & techniques designed and arranged and used for long-term bit preservation

Integrity check Coordination

Bit Repository

General System Layer

Pillar Layer

| *Pillar1* | *Pillar 2* | *Pillar 3* | *Pillar 4* | *Pillar 5* | *Pillar 6* | *Pillar 7* | *Pillar 8* |

...

Integrity – Bit error

Risk: Bits can change value
1. Error has occurred in Backup
2. File is corrupted
3. Error is not discovered
4. Cannot determine which file is intact

File in form
of bit stream
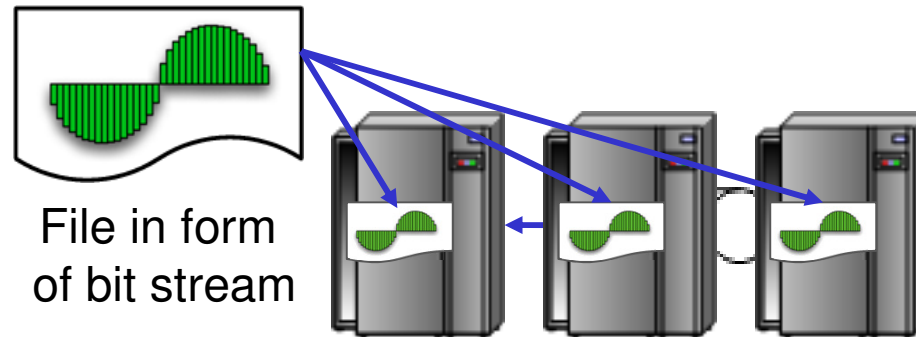
Solutions:
1. No backup. All are copies of data
2. Vote on which copy that is the right one

Bit error – System Layer

Bit Repository (BR)
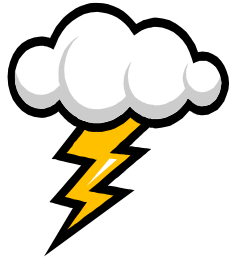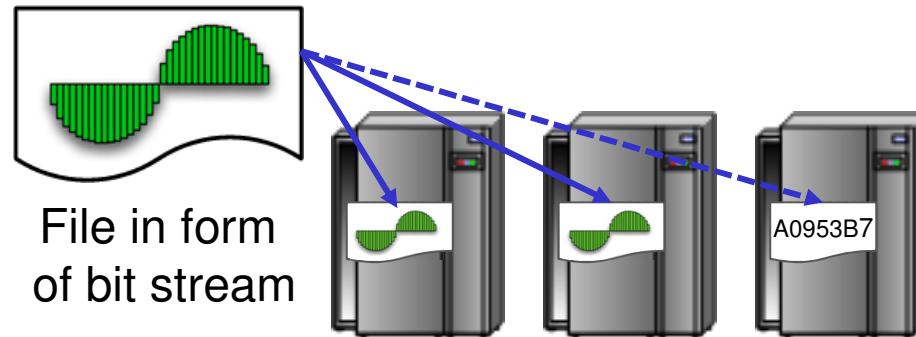
System Layer

Pillar Layer

...
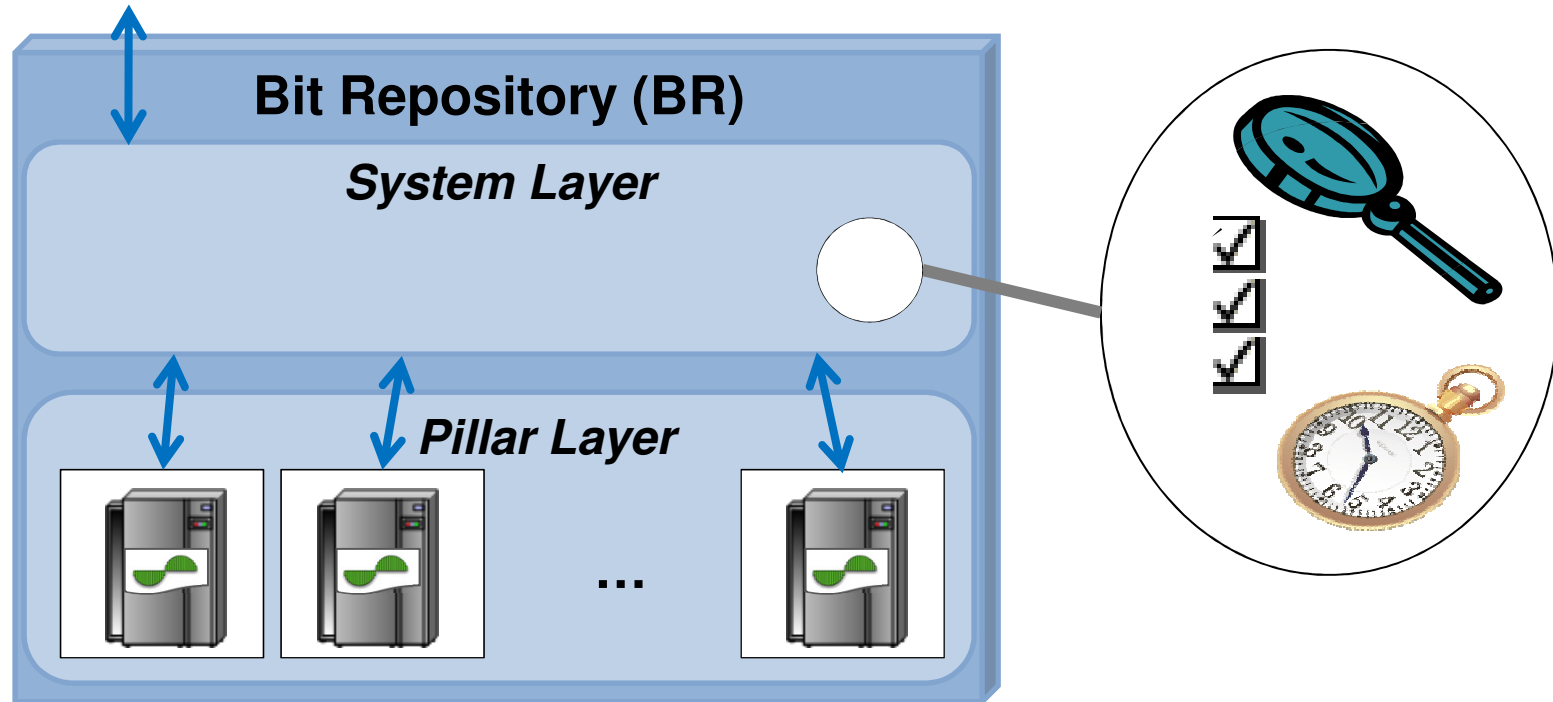
Solutions:

1. System layer checks and follow-up on basis of comparing copies
2. Minimum three voters, *optimize by checksums*

Integrity – Bit error

Risk: Bits can change value

1. Error has occurred in Backup
2. File is corrupted
3. Error is not discovered
4. Cannot determine which file is intact
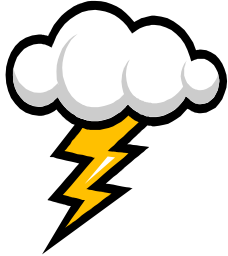
File in form
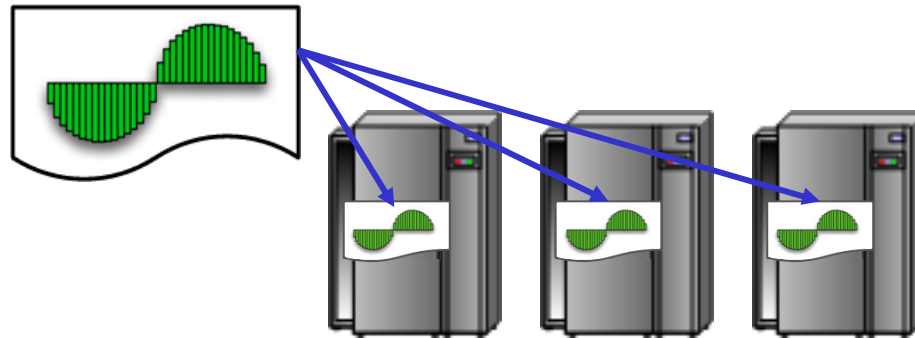of bit stream

A0953B7

Solutions:

1. No backup. All are copies of data
2. Vote on which copy that is the right one
3. Introduce checksums of files to discover errors
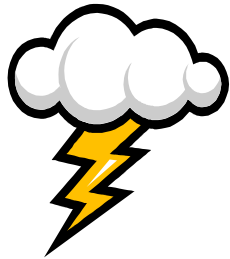
Integrity – Bit error

Risk: The same error occurs for more copies
1. Same hardware
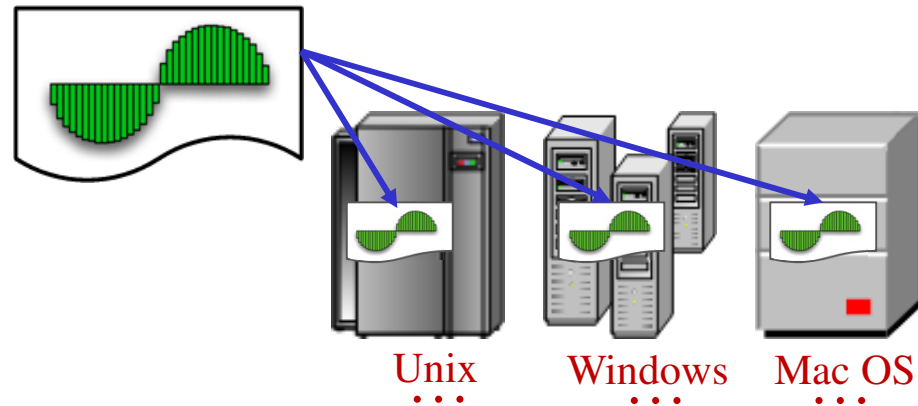2. Same software
3. Same vendor

Integrity – Bit error

Risk: The same error occurs for more copies
1. Same hardware
2. Same software
3. Same vendor
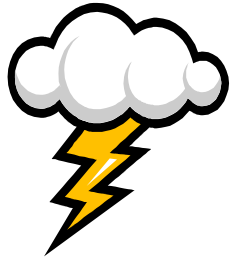
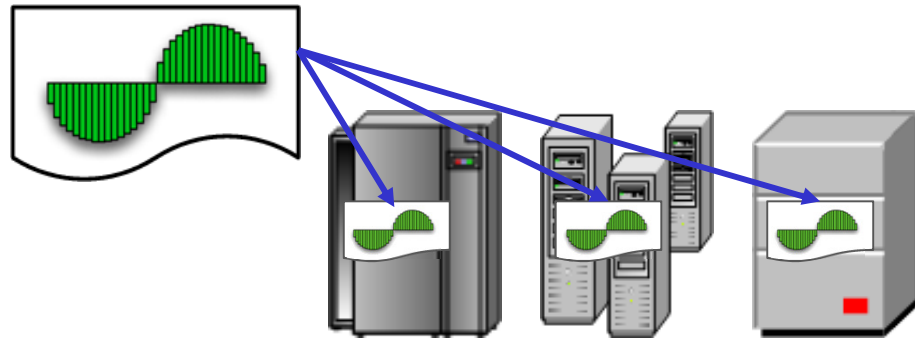Unix   Windows   Mac OS

Solutions:
1. Different hardware solutions
2. Different vendors
3. Different software (OS, interpreters, etc.)

Integrity – Disasters

Risk: All copies are damaged at the same time
1. Natural disasters
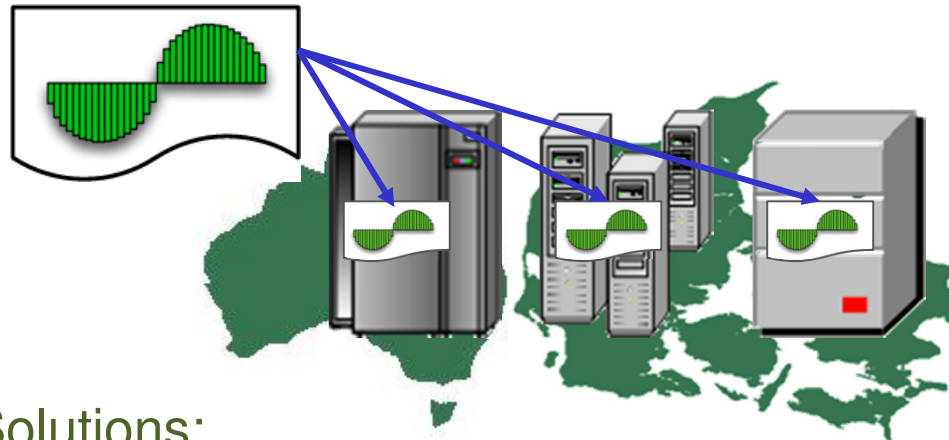2. Attacks in connection with war or terror

Integrity – Disasters

Risk: All copies are damaged at the same time
1. Natural disasters
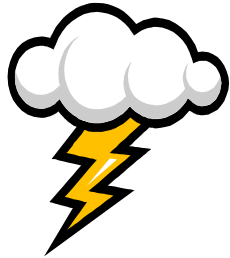2. Attacks in connection with war or terror

Solutions:
1. Different geographical locations

Integrity – Organisation

Risk: Errors/mistakes are made by the same person/org.
1. The same person has access and has delete rights
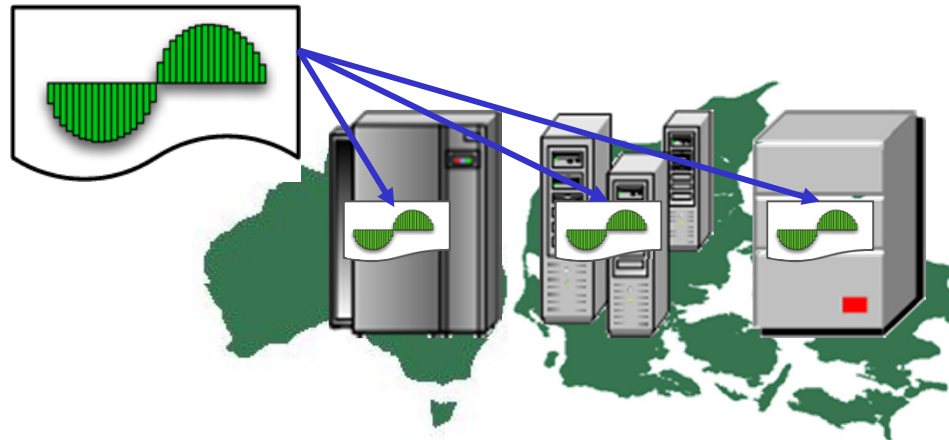2. The same person makes procedural mistakes

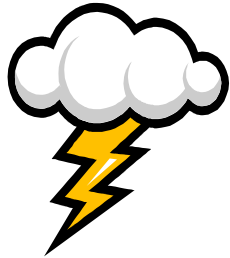# Integrity – Organisation

Risk: Errors/mistakes are made by the same person/org.
1. The same person has access and has delete rights
2. The same person makes procedural mistakes
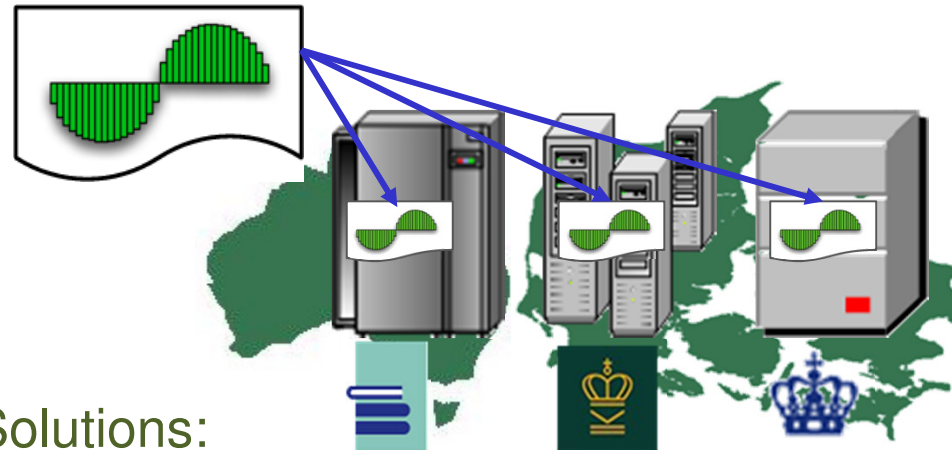
Solutions:
1. Different organisations

Confidentiality
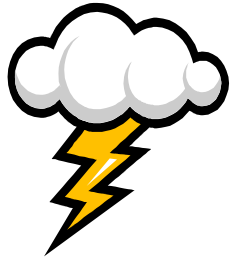
Risk: Unauthorised gets access to confidential data
1. Unauthorised gets access to Bit Repository
2. Unauthorised gets access to data from Bit Repository

Encryption can conflict integrity

A0953B7
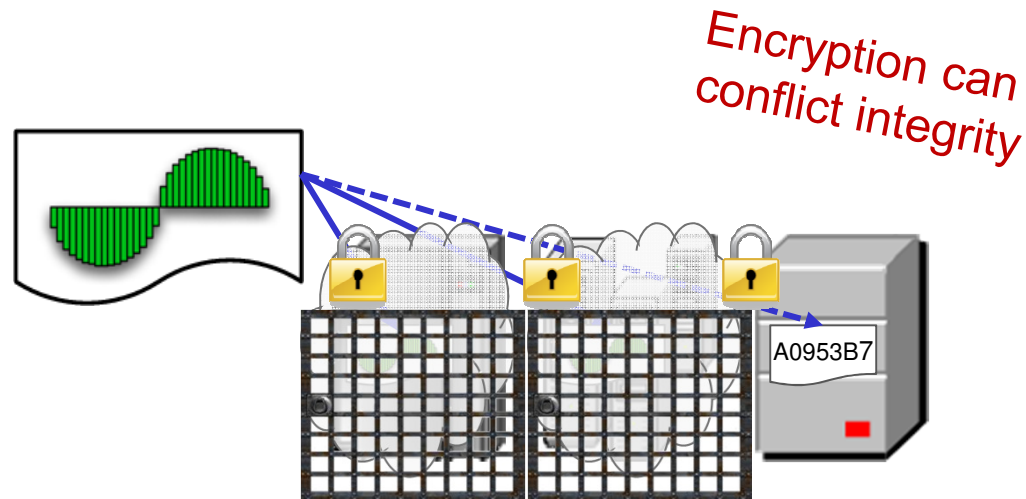
Solutions:
1. Authentication of users of pillars with copies
2. Encryption internally on pillar
3. Hardware secured in locked rooms

Confidentiality – System Layer

Bit Repository (BR)

System Layer

...

Pillar Layer

...

A09537
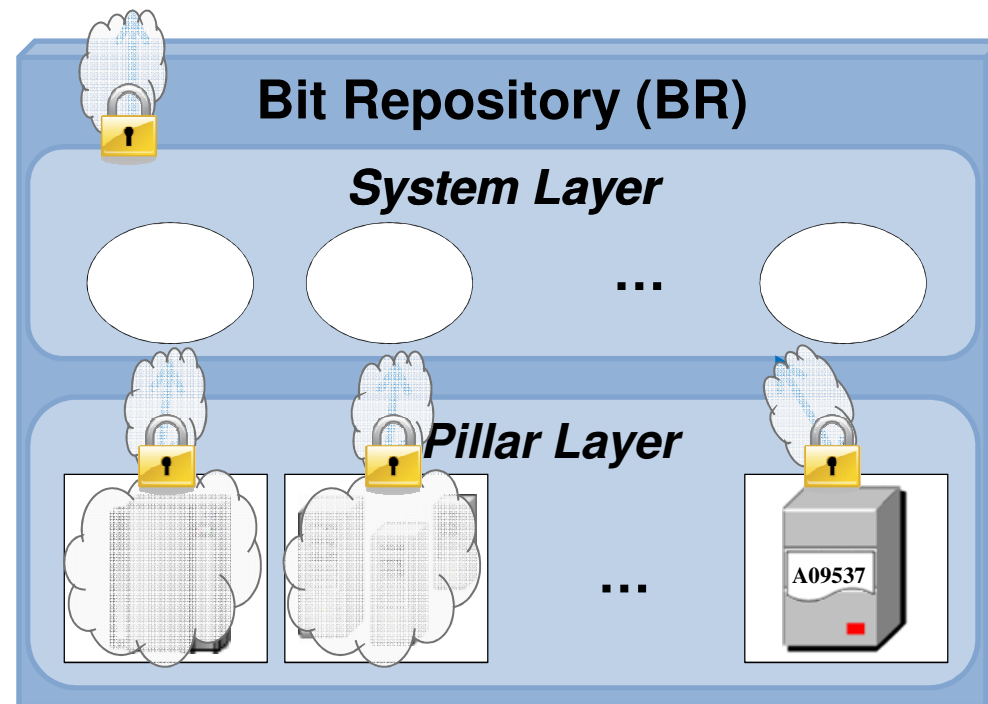
Likewise on System layer

Availability
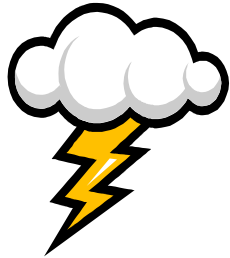
Risk: Cannot get access as required
1. Cannot get any response on request
2. Processing not possible in reality
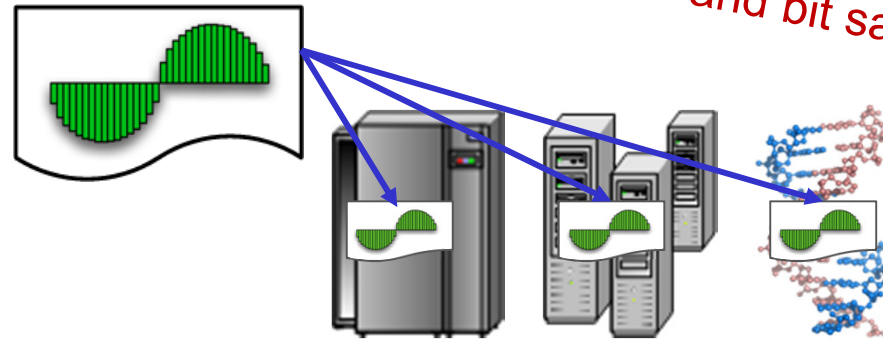
Processing can conflict confidentiality and bit safety

Solutions:
1. Specialised pillar with distributed architecture

Availability
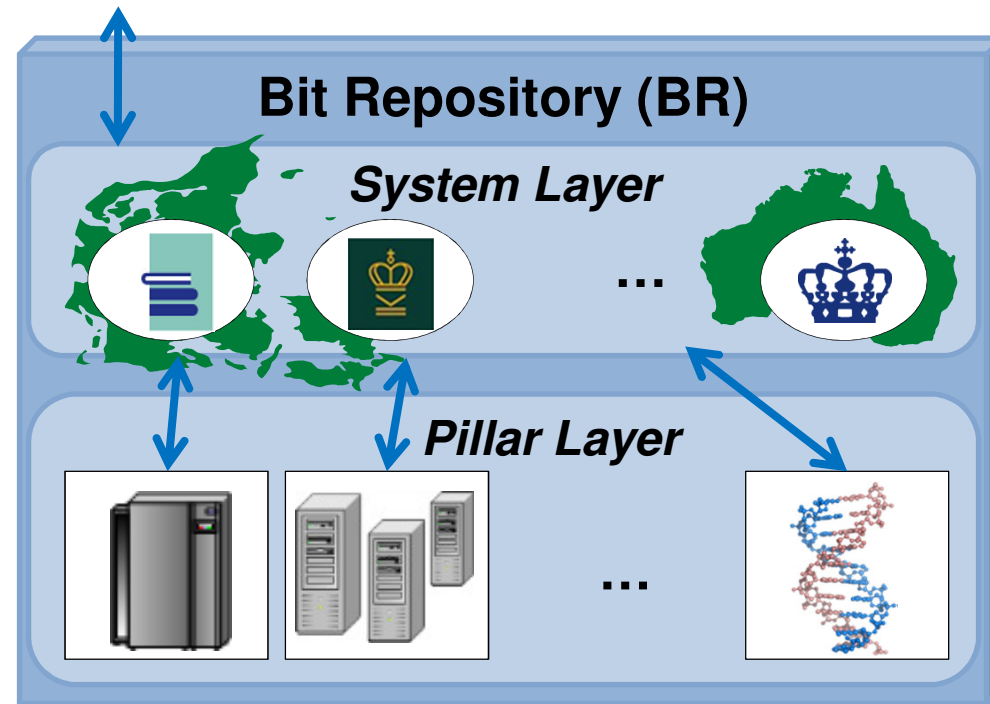
Solps: Solutions:
1. Redirection if access to a pillar is down
2. Distributed requests to different pillars
3. Scaling
4. Diversity,
5. …

Depends on what is required

Bit Repository Offering Solutions

BR

General System Layer

Pillar Layer

| Pillar 1 | Pillar 2 | Pillar 3 | Pillar 4 | Pillar 5 | Pillar 6 | Pillar 7 | Pillar 8 | Pillar 9 |

...
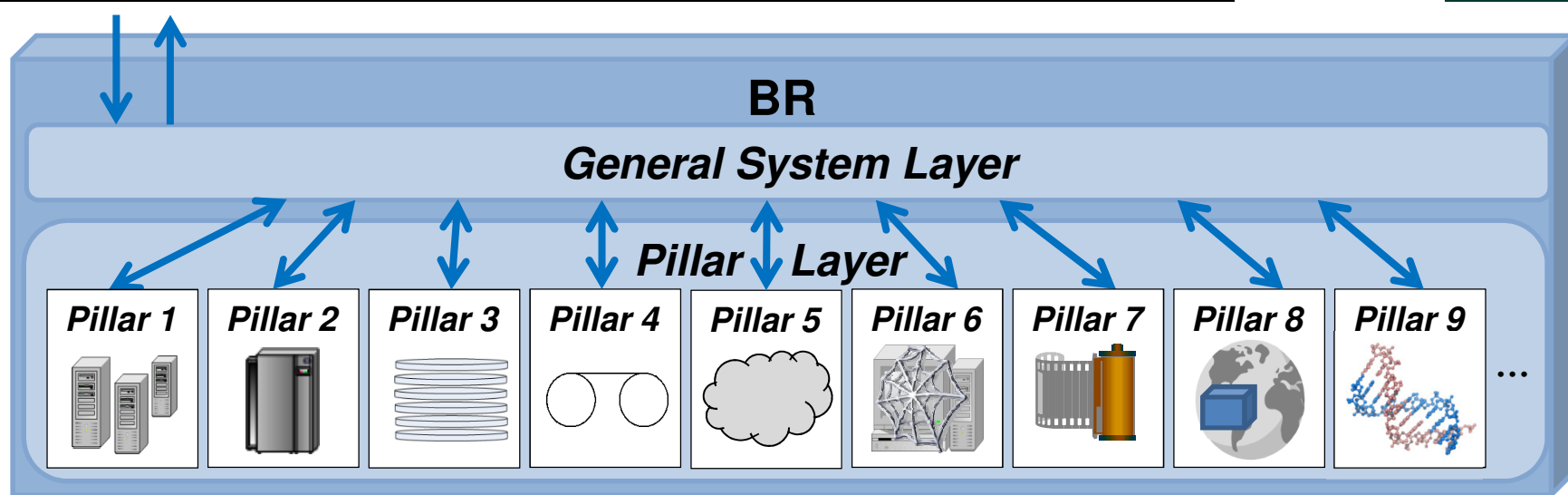
- Media
- Data safety
- Access speed
- On-line
- Off-line
- Organisational placement
- Geographical placement
- …

Bit safety   *Availability*   *Costs*

Confidentiality

Richard Gartner

Bit Safety

Value	Comment for preservationLevelType = BitSafety
Max	Maximum bit safety
VeryHigh	Very high bit safety
High	High bit safety
Medium	Medium bit safety
Low	Low bit safety
VeryLow	Very low bit safety
Min	Minimum bit safety

from http://id.kb.dk/vocabulary/

Bit Safety

Value	Comment for preservationLevelType = BitSafety
Max	**Maximum bit safety**
VeryHigh	Very high bit safety
High	High bit safety
Medium	Medium bit safety
Low	Low bit safety
VeryLow	Very low bit safety
Min	Minimum bit safety

Policy:
As high bit safety that we can get

Strategy 2013:
10 copies spread over 3 continents, both optical and magnetic medias, checked every …

Strategy 2050:
8 copies ; at lest 2 on Mars, at least two written to DNA, checked every …

Confidentiality

Value	Comment for preservationLevelType = Confidentiality
Max	Maximum confidentiality
VeryHigh	Very high confidentiality
High	High confidentiality
Medium	Medium confidentiality
Low	Low confidentiality
VeryLow	Very low confidentiality
Min	Minimum confidentiality

from http://id.kb.dk/vocabulary/

Confidentiality

Value	Comment for preservationLevelType = Confidentiality
Max	Maximum confidentiality
VeryHigh	Very high confidentiality
High	**High confidentiality**
Medium	Medium confidentiality
Low	Low confidentiality
VeryLow	Very low confidentiality
Min	Minimum confidentiality

Policy:
Only restricted access, where it is as hard as it can get for others when skipping encryption

Strategy 2013:
No more than 2 copies, that are secured on off-line medias …

Strategy 2050:
??? …

Availability

Value	Comment for preservationLevelType = Availability
Max	Maximum availability
VeryHigh	Very high availability
High	High availability
Medium	Medium availability
Low	Low availability
VeryLow	Very low availability
Min	Minimum availability

…

Logical Preservation

Value	Comment for preservationLevelType = logicalStrategy
Migration	Migration of digital material to keep data interpretable
Emulation	Emulation of digital material to keep data interpretable
Technical	Technology preservation to keep data interpretable

...

Angela had more

Preservation Level information

preservationLevelType	Comment
Bit safety	Bit preservation
Confidetiality	Bit preservation
Availability	Bit preservation
Logical Preservation Strategy	Logical Preservation
…	

- Policy
  - With institution preservation policies
  - Express values - Same over time
- Strategy
  - Requirements for fulfilment with current technologies …

Preservation Level in metadata

```xml
<digiprovMD CREATED="2013-01-18T19:28:01.456+01:00" ID="Premis1">
  <mdWrap MDTYPE="PREMIS">
   <xmlData>
     <preservationLevel xmlns:xlink="http://www.w3.org.1999/xlink" xsi:…">
      <preservationLevelValue>bitSafetyHigh</preservationLevelValue>
      <preservationLevelDateAssigned>2013-01-18T19:28:01.458+01:00
      </preservationLevelDateAssigned>
     </preservationLevel>
     <preservationLevel xmlns:xlink="http://www.w3.org.1999/xlink" xsi:…">
        <premis:preservationLevelValue>logicalStrategyMigration
        </premis:preservationLevelValue>
        <preservationLevelDateAssigned>2013-01-18T19:28:01.459+01:00
        </preservationLevelDateAssigned>
     </preservationLevel>
     <preservationLevel xmlns:xlink="http://www.w3.org.1999/xlink" xsi:…">
        <preservationLevelValue>confidentialityLow</preservationLevelValue>
        <preservationLevelDateAssigned>2013-01-18T19:28:01.460+01:00
        </preservationLevelDateAssigned>
     </preservationLevel>
   </xmlData>
  </mdWrap>
</digiprovMD>
```

Identification in the future

Angela and Robert said:
what is an intellectual entity?

As for representations

15AE9513

Producer → Object → **Service Provider** ← Object id. & ~~Service~~ ← **Con-sumer**

15AE9513

Producer ← Object id. ← **Service Provider** → Object → **Con-sumer**

Go back finding earlier versions (Richard Gartner)
Preservation messages (Huw Jones)

Functions (Eld)   Search rather than identification (tagging, …)?

Packaging – METS (Richard Gartner) -- WARC?

Fragments for commenting
(Carefull with preservation!!!!)

Preservation Level in metadata

```
<techMD CREATED="2013-01-18T19:28:01.426+01:00" ID="PremisObject1">
  <mdWrap MDTYPE="PREMIS:OBJECT">
   <xmlData>
    <object xmlns:xlink="http://www.w3.org.1999/xlink" xsi:…"…>
     <objectIdentifier>
      <objectIdentifierType>UUID</premis:objectIdentifierType>
      <objectIdentifierValue>41d153d0-0099-11e2-9397-005056887b67
      </objectIdentifierValue>
     </objectIdentifier>

     …
     <linkingIntellectualEntityIdentifier>
      <linkingIntellectualEntityIdentifierType>UUID
      </linkingIntellectualEntityIdentifierType>
      <linkingIntellectualEntityIdentifierValue>
        41d153d1-0099-11e2-9397-005056887b67
      </linkingIntellectualEntityIdentifierValue>
     </linkingIntellectualEntityIdentifier>
    </object>
   </xmlData>
  </mdWrap>
</techMD>
```
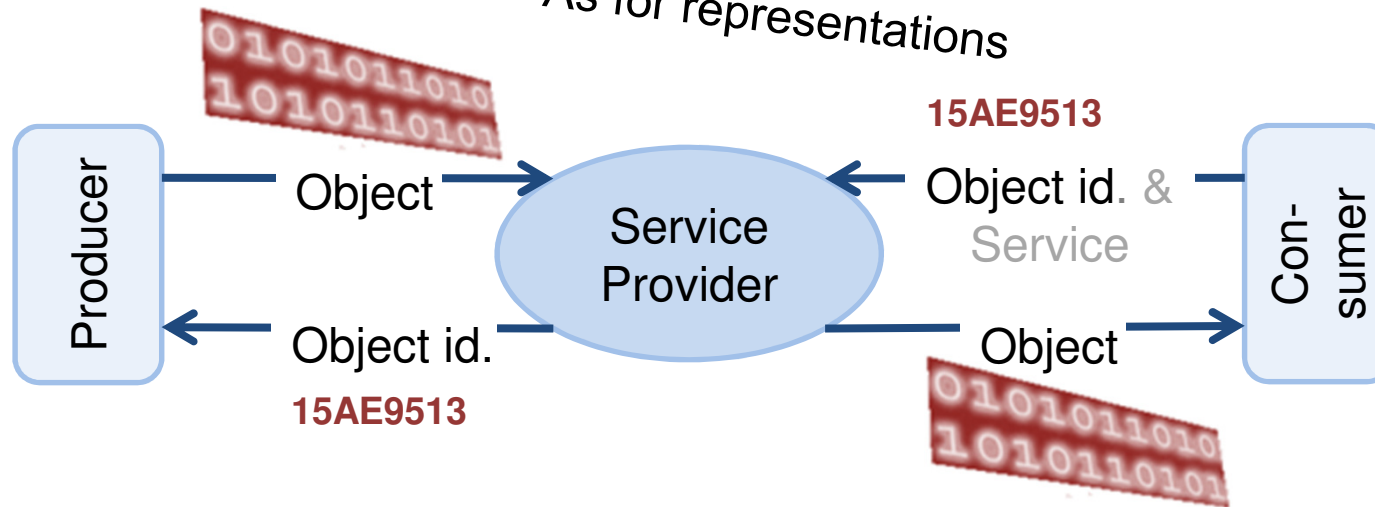
What is an intellectual entity?

Summary

PREMIS, METS and preservation metadata:

emerging trends and future directions

It depends

- "Choosing" preservation metadata standards
  - ◦ Which and how

    Community, sustainability, requirements

- Preserving preservation metadata

  - ◦ Which and how

    Share/advert

    Requirements

    Crowed sourcing

- Some challenges for the future

    Open linked data

  - ◦ Definition of **preservation levels** and **intellectual entities**
  - ◦ Expressing **preservation levels** and **intellectual entities**
  - ◦ Expressing **preservation levels** and **intellectual entities** over time

Questions and Comments