Regulations and recommendations about file formats

IASSIST 2022-06-09 Session G2 Data literacy

- Magnus Geber
- Benjamin Yousefi

Riksarkivet
Swedish National Archives

File formats for long time preservation (Magnus Geber)

Riksarkivet

Long time preservation of digital files

- Archival sector
- Long time preservation and access
- Migration strategy, conversion of files
- OIAS model, metadata
- Technically, set demands on file formats standardized, limited number

Riksarkivet

Format registries

PRONOM the primary format register

- Manage by British TNA

GDFR and UDFR

- Community base, did not succeed

Wikidata and NARA

- Mentioned as resources

Riksarkivet

Identifiers and Validators

Jhove

- Format identification, Validation, and Characterization tool

Droid and Fido

- Format identifications, related to Pronom

veraPDF, Jpylzer and MediaConch

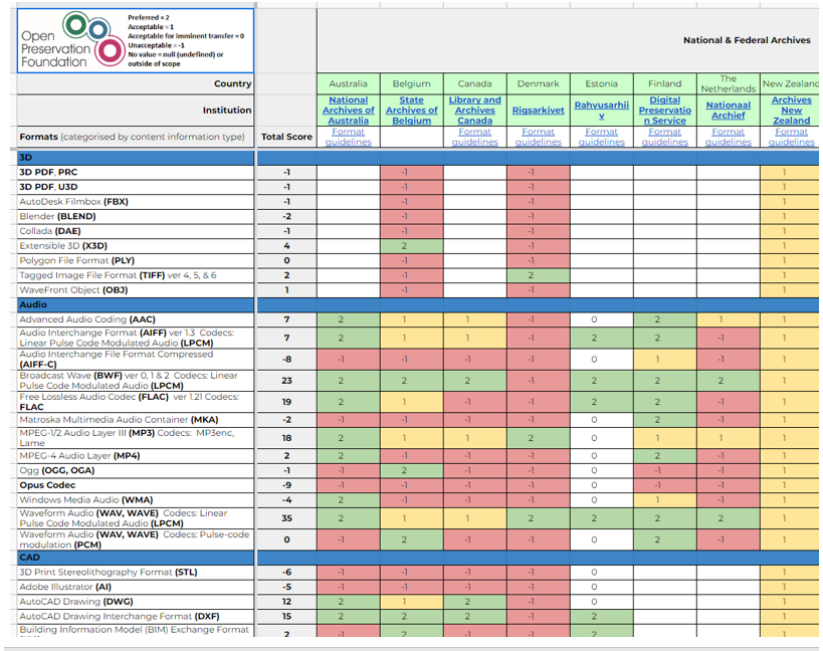
- Validators, format specific, for PDF, JP2 and AV (Matroska, LPCM, FFV1)

OPF (Open Preservation Foundation) manage many tools

Formats recommendations and regulations

International Comparison of Recommended File Formats

- Present recommendations from different culture heritage institutions
- Developed within OPF
- https://openpreservation.org/resources/member-groups/international-comparison-of-recommended-file-formats/



Open Preservation Foundation — Preferred = 2, Acceptable = 1, Acceptable for imminent transfer = 0, Unacceptable = -1, No value = null (undefined) or outside of scope		National & Federal Archives							
Country		Australia	Belgium	Canada	Denmark	Estonia	Finland	The Netherlands	New Zealand
Institution		National Archives of Australia	State Archives of Belgium	Library and Archives Canada	Rigsarkivet	Rahvusarhiiv	Digital Preservation Service	Nationaal Archief	Archives New Zealand
Formats (categorised by content information type)	Total Score	Format guidelines		Format guidelines	Format guidelines	Format guidelines	Format guidelines	Format guidelines	Format guidelines
3D									
3D PDF, PRC	-1		-1		-1				1
3D PDF, U3D	-1		-1		-1				1
AutoDesk Filmbox **(FBX)**	-1		-1		-1				1
Blender **(BLEND)**	-2		-1		-1				1
Collada **(DAE)**	-1		-1		-1				1
Extensible 3D **(X3D)**	4		2		-1				1
Polygon File Format **(PLY)**	0		-1		-1				1
Tagged Image File Format **(TIFF)** ver 4, 5, & 6	2		-1		2				1
WaveFront Object **(OBJ)**	1		-1		-1				1
Audio									
Advanced Audio Coding **(AAC)**	7	2	1	1	-1	0	2	1	1
Audio Interchange Format **(AIFF)** ver 1.3, Codecs: Linear Pulse Code Modulated Audio **(LPCM)**	7	2	1	1	-1	2	2	-1	1
Audio Interchange File Format: Compressed **(AIFF-C)**	-8	-1	-1	-1	-1	0	1	-1	1
Broadcast Wave **(BWF)** ver 0, 1 & 2, Codecs: Linear Pulse Code Modulated Audio **(LPCM)**	23	2	2	2	-1	2	2	2	1
Free Lossless Audio Codec **(FLAC)** ver 1.21 Codecs: **FLAC**	19	2	1	-1	-1	2	2	-1	1
Matroska Multimedia Audio Container **(MKA)**	-2	-1	-1	-1	-1	0	2	-1	1
MPEG-1/2 Audio Layer III **(MP3)** Codecs: MP3enc, Lame	18	2	1	1	2	0	1	1	1
MPEG-4 Audio Layer **(MP4)**	2	2	-1	-1	-1	0	2	-1	1
Ogg **(OGG, OGA)**	-1	-1	2	-1	-1	0	-1	-1	1
Opus Codec	-9	-1	-1	-1	-1	0	-1	-1	1
Windows Media Audio **(WMA)**	-4	2	-1	-1	-1	0	1	-1	1
Waveform Audio **(WAV, WAVE)** Codecs: Linear Pulse Code Modulated Audio **(LPCM)**	35	2	1	1	2	2	2	2	1
Waveform Audio **(WAV, WAVE)** Codecs: Pulse-code modulation **(PCM)**	0	-1	2	-1	-1	0	2	-1	1
CAD									
3D Print Stereolithography Format **(STL)**	-6	-1	-1	-1	-1	0			1
Adobe Illustrator **(AI)**	-5	-1	-1	-1	-1	0			1
AutoCAD Drawing **(DWG)**	12	2	1	2	-1	0			1
AutoCAD Drawing Interchange Format **(DXF)**	15	2	2	2	-1	2			1
Building Information Model (BIM) Exchange Format	2	-1	2	-1	-1	2			1

Riksarkivet

Selection of file formats

-Different criteria based on different models

Example from Danish National Archives

- Prevalence
- Lifespan
- Documentation
- Licensing
- Structure
- Significant properties
- Dissemination
- Searchability
- Interoperability
- Testing
- Compression
- Storage
- Migration
- Compatibility

Riksarkivet

- Proposed government agency regulations (Benjamin Yousefi)

Riksarkivet

FormatE – Two drafts

- “TeK” – technical requirements
  - Focus on the *technical condition, state* of an electronic act
- “ArK” – archival requirements
  - Focus on the *archival condition, state* of a public electronic act

FormatE – Overview

A separation of concern

- TeK – technical details
- ArK – the *form and function*

  to represent a content

Riksarkivet

FormatE – TeK and ArK

TeK

- What is required by materiel and methods to achieve a consistent coding and decoding of electronic acts
- What materiel and methods ***are appropriate*** for *durability*, *permanency*, (cf. constant, consistent)
- That is, *defines* but does not impose what to use

ArK

- Additional requirements imposed by the archival agency, technical or other,
  - what is required to understand a public electronic act
- What materiel and methods ***to use*** to produce public electronic acts that are appropriate for *archival*

FormatE – Materiel and methods

- Hardware and software, documentation and any other information necessary to code, decode an electronic act.
- For example, algorithms, articles, computer, formats, literature, program, source code, specifications.
- Usually electronic but not necessary, compare storage of digital information on optical, magnetic disks, or on paper (punched cards).

FormatE – Selection of specifications

- What are the criteria for
  - defining (TeK) appropriate technical requirements, and
  - imposing (ArK) archival requirements?

FormatE – Selection of specifications

TeK

- What is required to code, decode consistently over time
- *Objective criteria*
- That is, true for anyone that wants to code, decode

ArK

- What is required by the archival agency to handle, manage the public electronic acts?
  - Amongst other things, to code, decode
- *Subjective criteria*
- Depends on the organization's resources – present and planned for the future

FormatE – Selection of specifications

TeK

- Current draft, specifications that are common amongst archival agencies
- Technical requirements for those specifications or of already regulated specifications,
    - such as electronic invoices, signatures

ArK

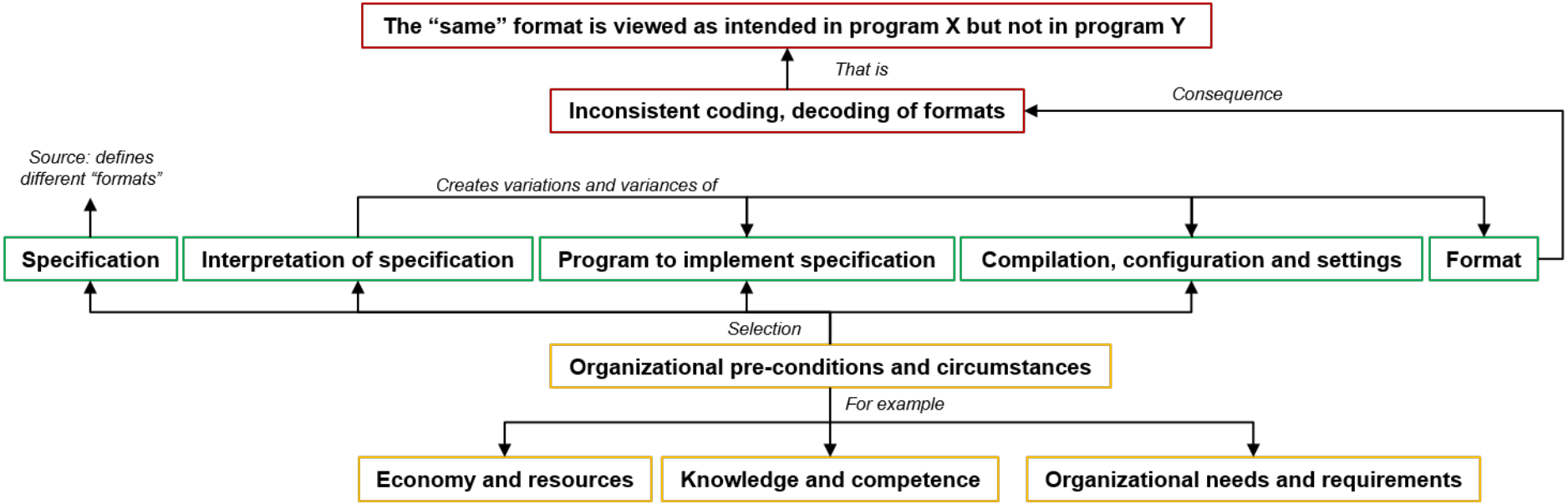
- Current draft, Riksarkivet:
- Knowledge and competence
- Budget and finances
- Established practices

Riksarkivet

FormatE – On “are appropriate for durability, permanency”

- For example,
  - encryption is not suitable for the long-term use, but
  - each public agency still has to decide whether or nor the use of encryption overweighs the need for preservation,
  - the agency’s archival agency could however require decryption before delivery.

FormatE – A simplified overview of the problem

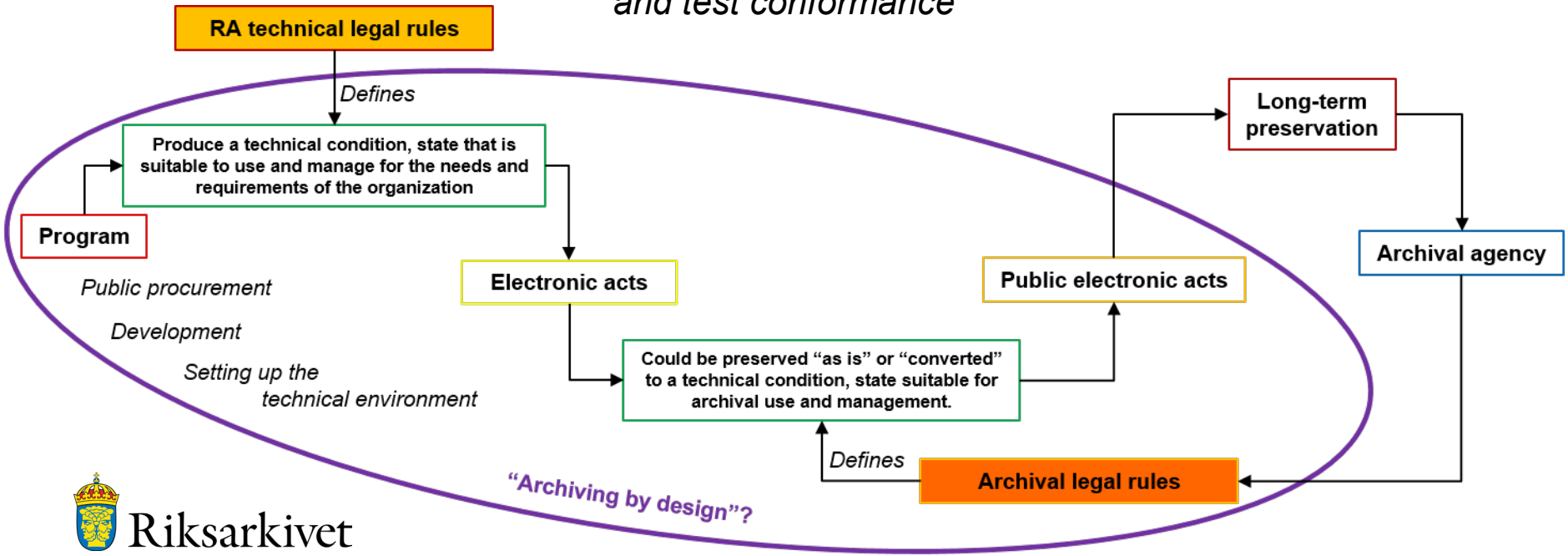


Riksarkivet

FormatE – A simplified overview of the regulation

A change in focus *– From "formats" to "specifications" and their "implementation" [formats]*

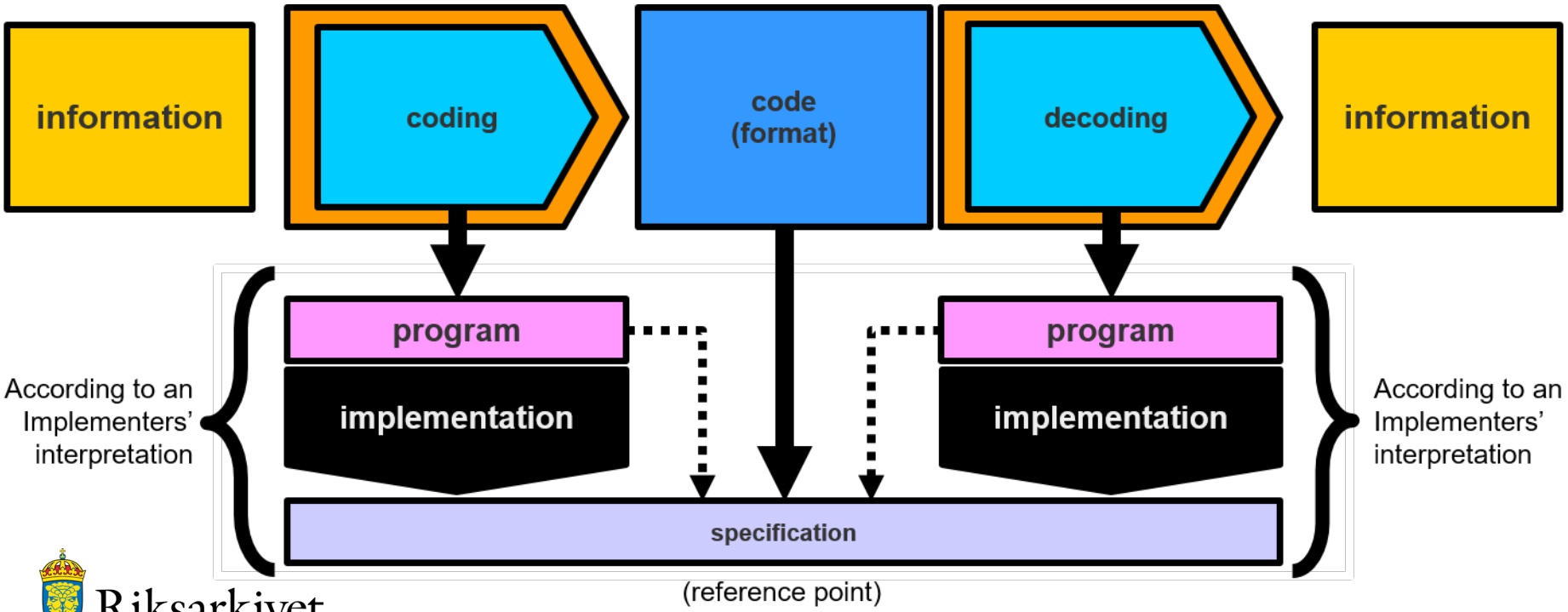
For example, "registries over specifications" and documentation on how to implement them and test conformance



Riksarkivet

FormatE – The model

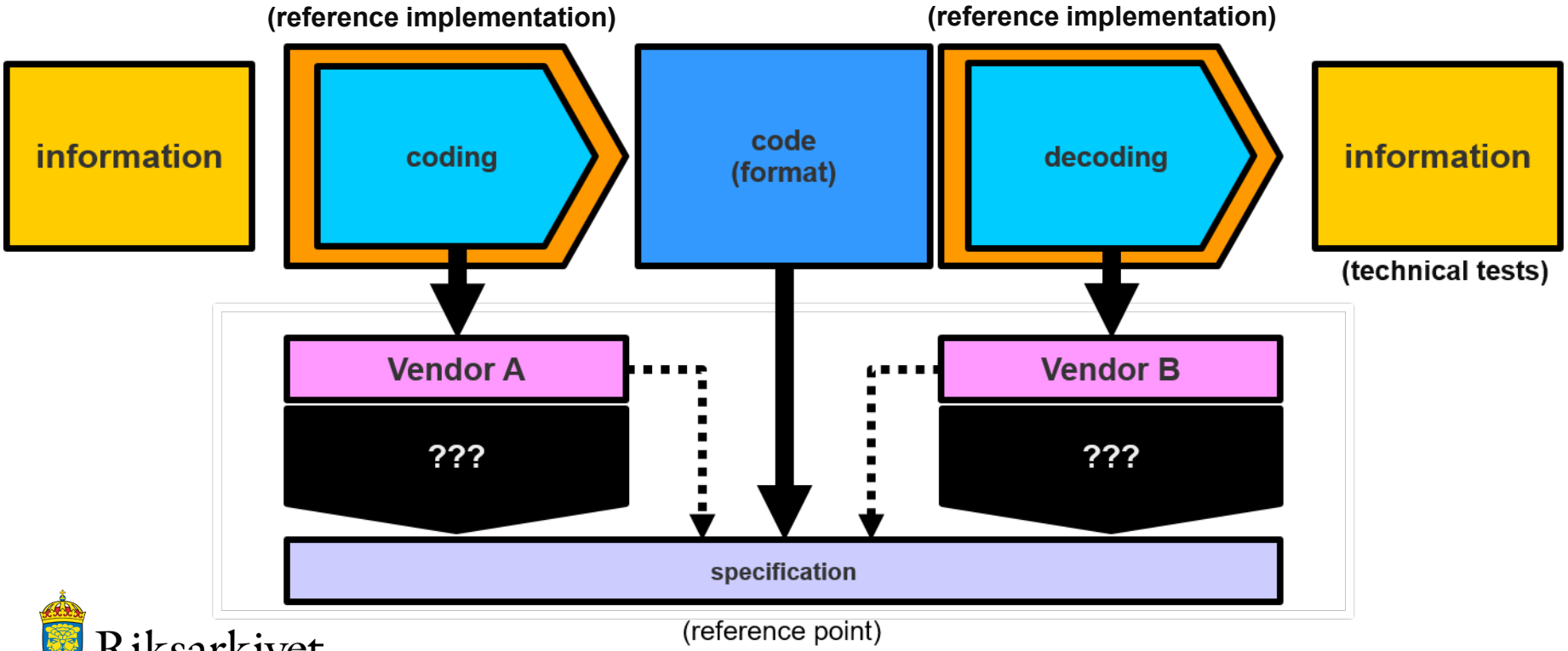
A change in focus – methods for measurability



Riksarkivet

FormatE – The model

A change in focus – methods for measurability



Riksarkivet

FormatE – Terminology

Swedish	English
Handling	*Act*
Allmän handling	*Public act*
Elektronisk handling	*Electronic act*
Allmän elektronisk handling	*Public electronic act*
Framställa	*Produce, manufacture*
Beständighet	*Durability, permanency, (cf. constant)*
Arkivbeständighet	*Archival permanence*
Arkivrättsliga krav	*Archival regulatory requirements*
Tekniska krav	*Technical requirements*
Tekniskt skick	*Technical condition, state [of the electronic act]*

Riksarkivet

Presentation is terminated

Riksarkivet