UNIVERSITY OF SKÖVDE

# Sitter vi på en datasophög – Vad är ett beständigt dataformat?

Bitr. Professor Björn Lundell, Ph.D.
**S**oftware **S**ystems **R**esearch **G**roup
Högskolan i Skövde
**bjorn.lundell@his.se**

@

Forskningsdataseminarium
30 maj 2017,
BTH, Karlskrona

UNIVERSITY OF SKÖVDE

# *Agenda …*

- Introduktion
- **Filformat** och dess relation till standarder
- **Utgångspunkter** och **utmaningar**
- Teknisk, ekonomisk och juridisk **inlåsning i filformat**
- **Öppna** och slutna **filformat** (som används av forskare)
- **Strategier** för IT-upphandling **från tidigare studie** med särskild relevans för **arkivering av forskning**
- **Öppen programvara** och dess roll för implementation av filformat för långsiktigt digitalt bevarande
- **Sammanfattning** – några **öppna frågor**

UNIVERSITY OF SKÖVDE

# Vad är ett **filformat**?

- "A file format is a **metho**d of **storing digital information in a computer file**, allowing its later use by computer systems or people. There are **thousands of different file formats** for different kinds of digital content and there may be **several different versions** of the „same" file format. A file format is often confused with the software most commonly used to create or use it."

  (The National Archives, 2011)

- OpenDocument 1.2 (ODF, Open Document Format for Office Applications) är ett exempel på ett filformat för dokument som definieras av OASIS (Organization for the Advancement of Structured Information Standards).

UNIVERSITY OF SKÖVDE

# *Enkla och (sammansatta) containerformat ...*

- “File formats are specific patterns or structures that organize and define data. Some formats contain only one stream of uncompressed data, others may contain codecs to encode and compress the data and others may support several streams of media. In addition to file formats, there are also container or encapsulating formats. These formats can contain and support various types or layers of data and metadata. Each of these formats may be handled by different programs, processes, or hardware but for the data stream to be interpreted properly, the information must be wrapped together.”

  (Library and Archives Canada, 2015)

- MXF (Material eXchange Format) är ett exempel på ett containerformat för professionell AV (video och ljudmedia) som definieras av SMPTE (Society of Motion Picture and Television Engineers)

UNIVERSITY OF SKÖVDE

# Standard?

- "... a published document that contains a **technical specification** or other precise criteria designed to be used consistently as a rule, guideline, or definition."

  (British Standards Institution: BSI, 2015)

- "… a publicly available **definitive specification** of procedures, rules and requirements, issued by a legitimated and recognized authority through voluntary consensus building observing due process, that establishes the baseline of a common understanding of what a given system or service should offer."

  (Jakobs, 2000, p. 17)

# Öppna standarder

*(SOU 2009:86 & EIF v1.0 & Kammarkollegiets ramavtal)*

- "The standard has been published and the **standard specification document** is available either **freely** or at a **nominal charge**. It must be permissible to all to copy, distribute and use it for no fee or at a nominal fee.
- The standard is adopted and will be maintained by a not-for-profit organisation, and its ongoing development occurs on the basis of an **open decision-making procedure** available to all interested parties (consensus or majority decision etc.).
- The intellectual property - i.e. **patents possibly present** - of (parts of) the standard is made **irrevocably available** on a **royalty-free basis**.
- There are no constraints on the re-use of the standard."

(Kammarkollegiet, 2015a, 2015b, 2016)

UNIVERSITY OF SKÖVDE

# *Utgångspunkter och Utmaningar …*

- Organisationer använder många **olika programvaror**
- Organisationer hanterar filer i många **olika filformat**
- Organisationer behöver ofta förvalta och **modifiera** sin **programvara** och sina **filer** (digitala handlingar) under **mer än 30 år**, för offentliga organisationer betydligt längre
- Förvaltning och supportkontrakt för proprietär programvara tillhandahålls (upp till) 10 år
- **Programvara** som används för att skapa filer **kommer inte att vara tillgänglig** under hela den tid som filen behöver förvaltas och arkiveras (exempelvis i e-arkiv)

***… detta är några av de utmaningar som vi analyserat i ett antal forskningsprojekt genom åren ...***

UNIVERSITY OF SKÖVDE

# **Filer från forskning** måste arkiveras *...*

## *... men **kommer filformaten kunna tolkas** &*

## *… finns programvaror och leverantörer kvar?*

"Even if Google and Apple don't exist in the year 3015, how can we ensure our files, photos, movies, and more, can be preserved for ages to come? How can we future-proof our content for generations way down the line?"

(Smith, 2015)


Av liftarn (originally posted to Flickr as ABC80) [CC BY-SA 2.0 (http://creativecommons.org/licenses/by-sa/2.0)], via Wikimedia Commons https://upload.wikimedia.org/wikipedia/commons/a/af/ABC80.jpg

# Kan arkiven förvalta digitala fotografier och dokument (och alla andra typer av data)?

- "Vint Cerf, a 'father of the internet', says he is worried that all the images and documents we have been saving on computers will eventually be lost."

  (Ghosh, 2015)

- "Cerf recently spoke about this topic at the annual conference of the American Association of the Advancement of Science, warning that if we don't move now, we risk losing all the data we've created in the 21st century."

  (Smith, 2015)

*… vad är **risken för förlust av data och resultat från forskning?***

*… **kan filer konverteras till 'bättre' format utan förlust av data?***

UNIVERSITY OF SKÖVDE

# Finns det möjlighet att konvertera filer mellan olika format inför en arkivering?

- “Not all applications maintain backward compatibility with their own versions, to say nothing of ability to convert into and from a wide range of formats other than their own.”
  (Cerf, 2010, p. 31)

- Analys av 16106 doktorsavhandlingar från svenska lärosäten (publicerade på SwePub) visar att en minoritet av dessa PDF-filer (<10 %) uppfyller grundläggande krav på filformat för långsiktigt digitalt bevarande (PDF/A-1b).
  (Fischer & Lundell, 2013)

- Ännu opublicerade resultat visar att förlag genererar filer som inte uppfyller krav på långsiktigt digitalt bevarande: publicerade PDF-filer uppfyller inte kraven från PDF/A-1b.
  (Acknowledgement till Thomas Fischer för teknisk analys i pågående studie)

# *Utgångspunkter: En **ohållbar** digital infrastruktur ...*
# Varför används **standarder** som skapar **inlåsning**?

- Many public sector organisations conduct projects that “include requirements with reference to specific proprietary products, specific trademarks, specific technologies (controlled by a single company) and specific closed standards.”

  *–Björn Lundell, svar öppen konsultation till (draft) EIFv3*

- ”Utifrån ett konkurrensperspektiv är det ofta problematiskt när offentliga aktörer genomför IT-upphandlingar och ställer krav på slutna standarder”

  *–Dan Sjöblom, Generaldirektör, Konkurrensverket, förord till* (Lundell et al., 2016)

- “**FRAND** licenses create **barriers** for Open Source projects”

  (EC COM(2013) 455 final & SWD(2013) 224 final)

# *Utgångspunkter: En **hållbar** digital infrastruktur …*
## Kan inlåsning undvikas & innovation stimuleras?

- Digitala artefakter överlever proprietär programvara i alla scenarier
- En vägledning för öppna standarder och öppna filformat har publicerats av Kammarkollegiet

  joinup.ec.europa.eu/community/osor/news/sweden-updates-list-mandatory-it-standards
- Strategier för en hållbar digital infrastruktur som undviker inlåsning och stimulerar innovation är publicerad av Konkurrensverket

  www.konkurrensverket.se/nyheter/problem-med-slutna-standarder-vid-it-upphandlingar/

  www.konkurrensverket.se/nyhetsbrevsartiklar/it-standarder-inlasning-och-konkurrens/
- ”För att möjliggöra interoperabilitet och långsiktigt digitalt bevarande, använd endast öppna standarder och öppna filformat som har implementerats i programvara och som därmed är möjliga att tillhandahålla samt distribuera under olika licenser, inklusive alla licenser för öppen programvara.” (Lundell et al., 2016, s. 8)

UNIVERSITY OF SKÖVDE

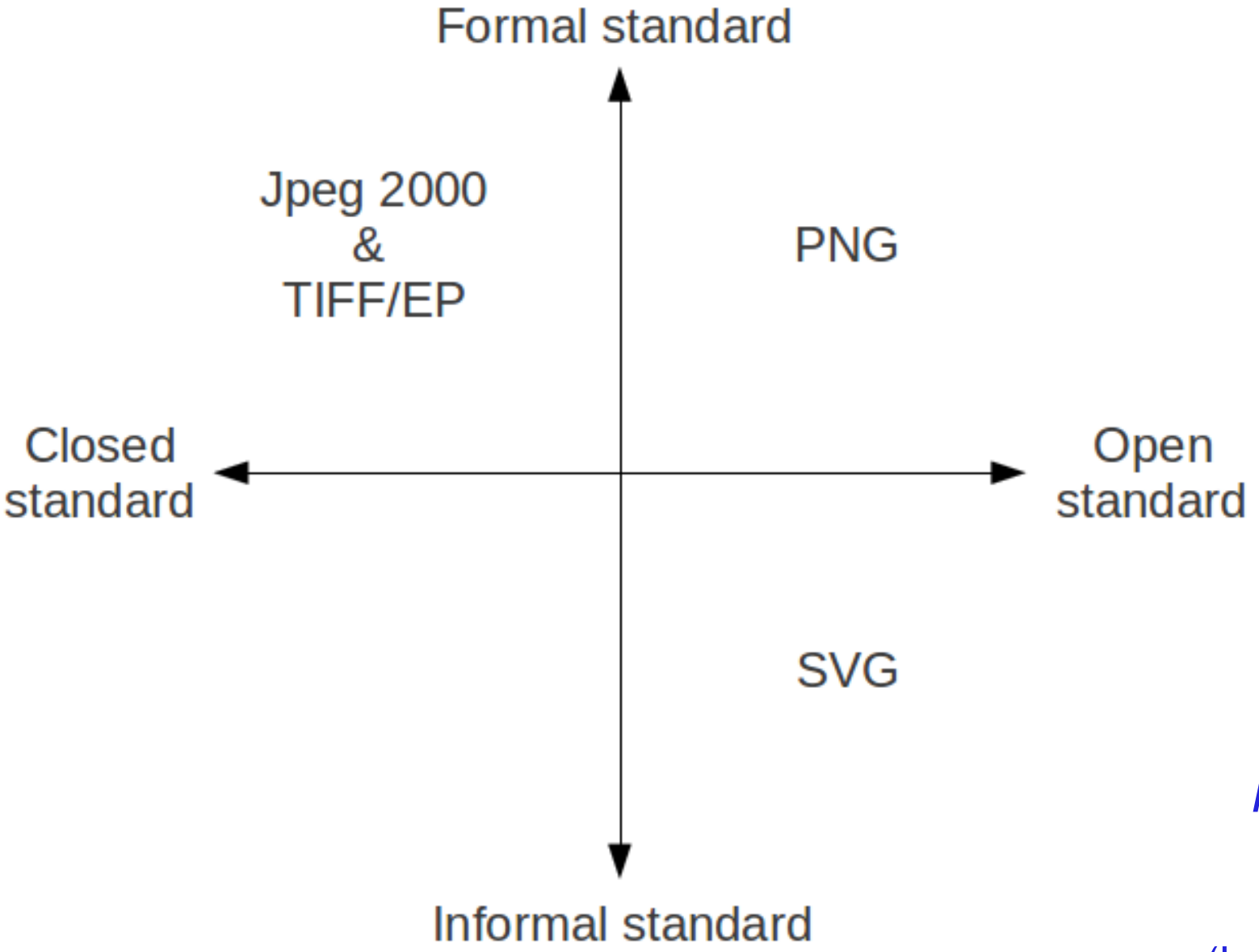
# *Utgångspunkter: En gigantisk utmaning ...*
## Teknisk, ekonomisk och juridisk **inlåsning** i **filformat**?

***Är det** nu och i framtiden …*

- **tekniskt** – kan filer tolkas korrekt?,
- **ekonomiskt** – kan kostnaden hanteras? *och*
- **juridiskt** – kan patentlicenser införskaffas?

*… **möjligt** att hantera och migrera filer från en inlåst situation så att det är **möjligt** att **tolka** samt **återanvända** arkiverade **data och resultat från forskning** (data, digitala handlingar samt program)?*

# *Öppna (och slutna) standarder* vs. *Formella (och informella) standarder ...*




*Långsiktigt digitalt bevarande* ***förutsätter*** *öppna standarder*

*Formella standarder är* ***inte*** *nödvändigtvis öppna*

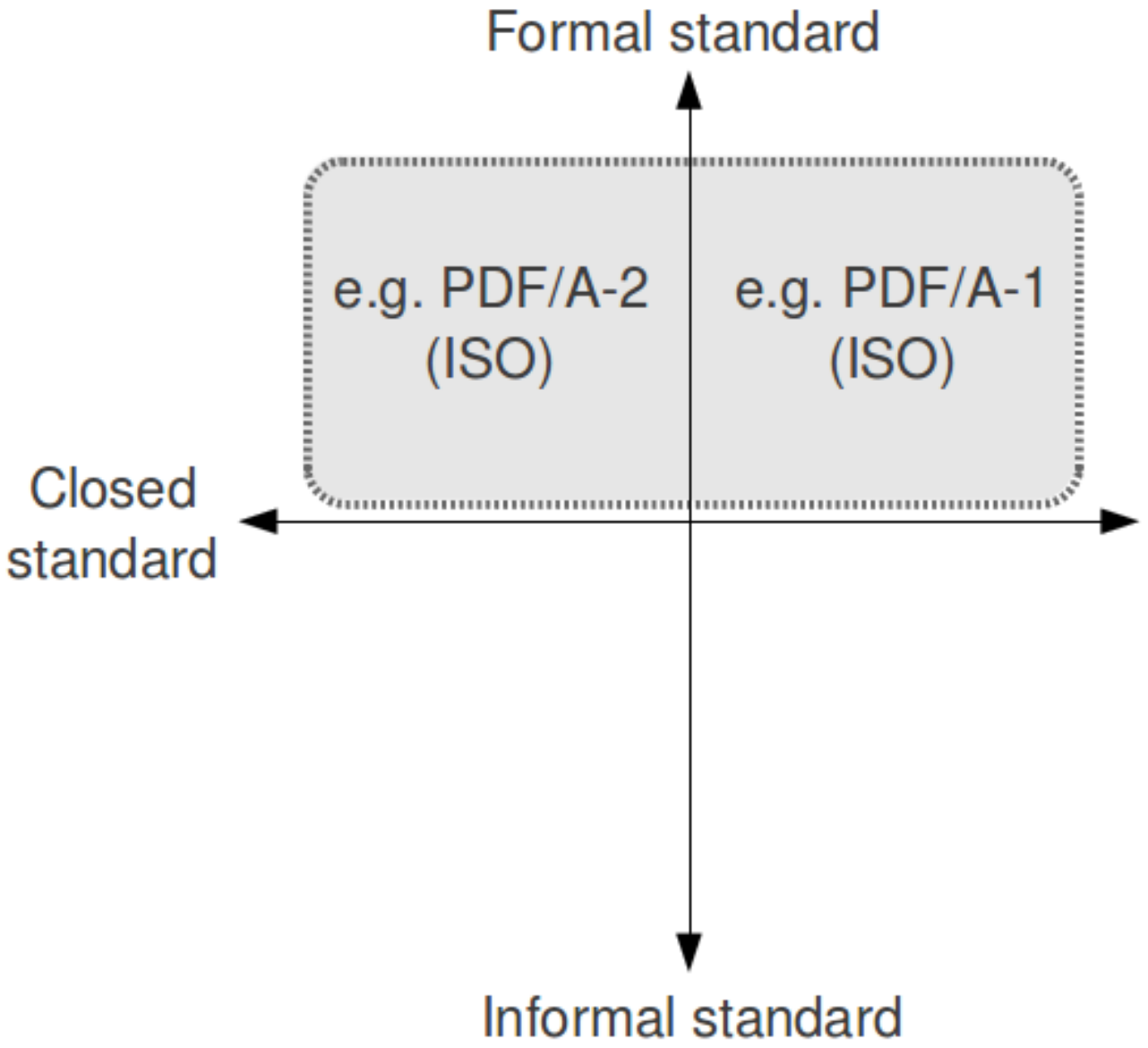
(Lundell et al., 2015)

# *Studie av standarder visar att endast öppna standarder kan implementeras i programvara ...*

- Vissa standarder förvaltas av fler än en organisation (exempelvis förvaltas PNG av både W3C och ISO)
- Vissa formella standarder (exvis PNG) kan implementeras i programvara som tillhandahålls under olika proprietära och olika open source-licenser (inga kända problem med patent för PNG-formatet)
- Olika patentdatabaser för en given standard innehåller **inkonsistent information** (exvis är innehållet för JPEG 2000 olika i ISO:s och ITU-T:s patentdatabaser)
- För vissa formella standarder (exvis JPEG 2000) är det **inte möjligt att komma i kontakt** med alla organisationer som deklarerat innehav av patent för att erhålla alla nödvändiga patentlicenser (och för vissa som det är möjligt att komma i kontakt med finns ovilja att erbjuda patentlicenser)

(Lundell et al., 2015)

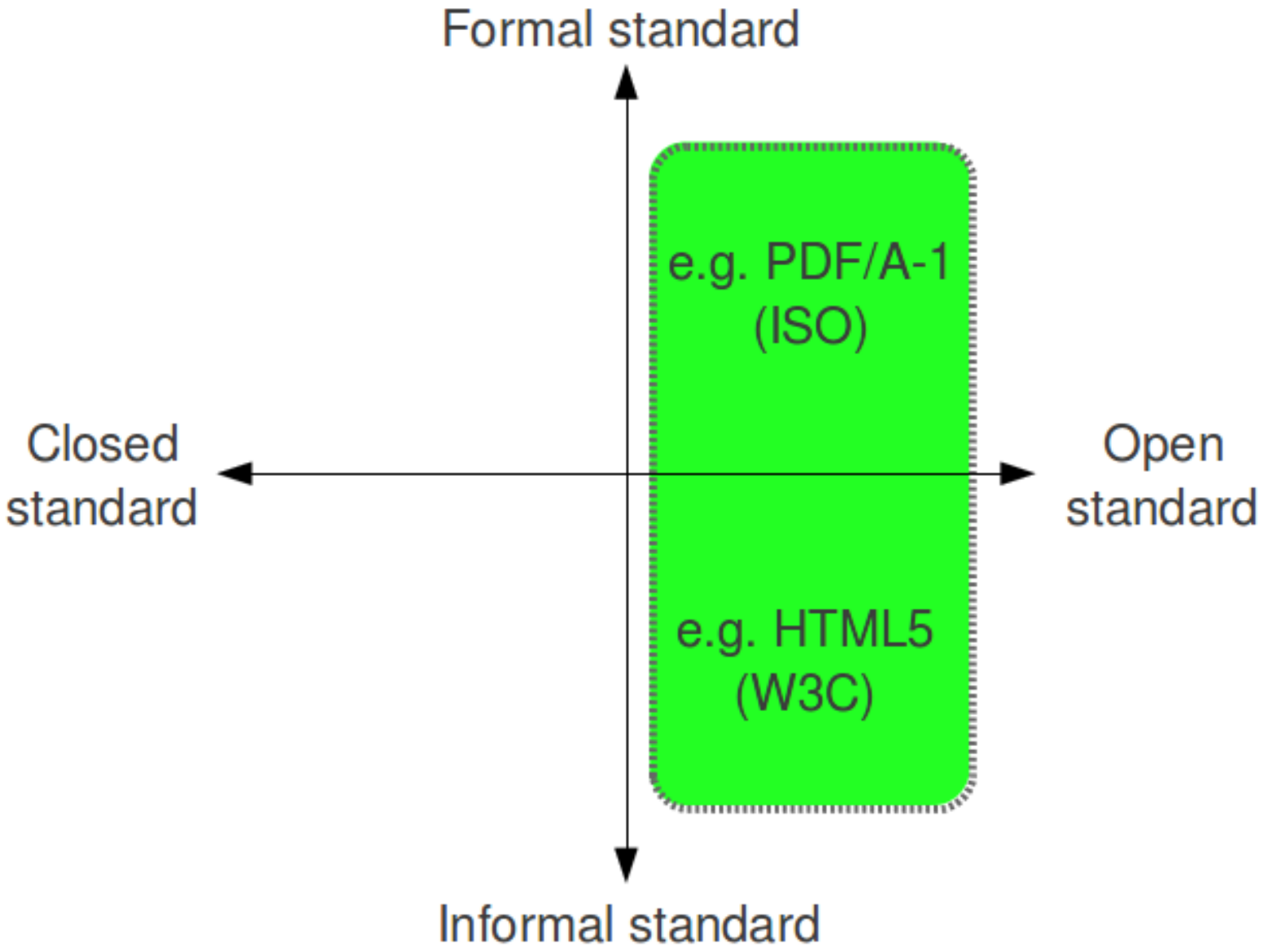
# *Implementation av patentbelastade formella standarder kräver **patentlicenser** för alla SEPs …*




*Kan alla patentlicenser för alla **SEPs** införskaffas för alla formella standarder?*

*(SEP, 'standard-essential-patent')*

# *Implementation av en **öppen standard** kan förväntas vara **oproblematisk** ...*




*Är alla relevanta standarder inkluderade i vägledningen (till SIC:s ramavtal 'Programvaror och tjänster 2014') som redovisar öppna standarder?*

# *Tidigare studie – med relevans för forskningen ...*
## *… ”**IT-standarder**, **inlåsning** och **konkurrens**”*

- Vilka strategier används av (olika typer av) organisationer för att **undvika** att påverkas av skadliga (konkurrensbegränsande) **inlåsningseffekter**?
- Vilka strategier används av (olika typer av) organisationer för att (utifrån en situation där en organisation redan sitter fast i en konkurrensbegränsande inlåsning) **komma ur** (’unlocking’) en **inlåsning**?
- Vilka (positiva och negativa) effekter för konkurrens på marknaden blir **konsekvensen av att referera till olika typer av IT-standarder** som förvaltas och tillhandahålls (av formella och informella standardiseringsorganisationer) under olika villkor?

(Lundell et al., 2016)

UNIVERSITY OF SKÖVDE

# *Två (av totalt sju) strategier …*
## *… med **särskild relevans för arkivering av forskning** …*

- “För att möjliggöra interoperabilitet och långsiktigt digitalt bevarande, använd endast öppna standarder och **öppna filformat** som har implementerats i programvara och som därmed är möjliga att tillhandahålla samt **distribuera under** olika licenser, inklusive **alla licenser för öppen programvara**.”

(Lundell et al., 2016, s. 8)

- “För att hantera data och handlingar som inkommer till en myndighet i slutna filformat, införskaffa innan upphandling alla nödvändiga rättigheter, inklusive **alla nödvändiga patentlicenser**, för dessa slutna filformat så att de kan implementeras i programvara som kan användas och distribueras under olika licenser, inklusive **alla licenser för öppen programvara**.”

(Lundell et al., 2016, s. 9)

# ***Frågeställningar om IT-upphandling*** *som* ***...***

## *… påverkar* ***filformat*** *och långsiktigt digitalt* ***bevarande***

- Hur kan **filformat** och processer för att införskaffa IT (däribland IT-upphandling) **skapa inlåsning** som hindrar långsiktig förvaltning och arkivering av de digitala artefakter (data, handlingar, program, etc.) som forskare utvecklar och hanterar i olika projekt?
- Hur & varför skapar forskare och myndigheter **inlåsning** som **hindrar långsiktig förvaltning av data och resultat från forskning**?
- Hur många forskare använder långsiktigt förvaltningsbar programvara som genererar data i långsiktigt förvaltningsbara öppna filformat?
- Hur många myndigheter ställer krav på att programvara ska kunna arkiveras redan vid IT-upphandlingen?

# *Molninlåsning & transformationsinlåsning …*

- ”Vid införskaffande av en kostnadsfri molnlösning kan det, framgent, uppstå ett beroende – en **molninlåsning** – av en enskild lösning som kan vara tekniskt, pedagogiskt och legalt komplicerad att ta sig ur med bibehållande av alla data och meta-data för vidare förvaltning och drift i en annan miljö. Ibland anförs kostnadseffektivitet som skäl för att införa molnlösningar i skolan, samtidigt som det kan konstateras att tidigare analyser i Sverige inte identifierat (och därmed ej heller kan redovisa) någon erfarenhet från någon organisation som gjort exit från en molnlösning.”
- Då “representation av data i ett ’gammalt’ befintligt system har en struktur som är okänd för den som ska utveckla en programvara och system för att **transformera data** innebär detta en **avsevärd risk för att inte korrekt kunna genomföra en transformation** till den struktur som behövs för det nya systemet.” (Lundell et al., 2016)

# ***Öppen programvara*** *(Open Source Software) ...*

- ... är programvara som tillgängliggjorts under en programvarulicens vilken godkänts av **Open Source Initiative** (OSI), se: www.opensource.org

- **Öppna standarder** enligt EIFv1 (d.v.s. enligt den definition som används i Kammarkollegiets ramavtal) **kan implementeras under olika licenser för sluten och under olika licenser för öppen programvara**, inklusive GPL (GNU General Public License version 3, se opensource.org/licenses/GPL-3.0)

UNIVERSITY OF SKÖVDE

# *Referensimplementation och öppen referensimplementation ...*

- "A **Reference Implementation** is an implementation of a specification which can be used as a **definitive interpretation** of the standard's specification."

(Lundell et al., 2012)

- "an **Open Source Reference Implementation** is a reference implementation of the specification of a standard that is licensed under an Open Source license"

(Lundell et al., 2012)

# *Specifikationer av standarder är ofullständiga och implementeras olika i olika programvaror ...*

*En teknisk specifikation av en standard är ofta ofullständig och implementeras ofta på olika sätt i olika programvaror ...*

> "For most software standards the **formal specification** is **insufficient** and the actual standard may **differ from across implementations**. ... the formal specification is **inherently incomplete** and the actual standard is **defined both through** the **written specification and through actual implementations**"
>
> (FLOSSPOLS, 2005)

UNIVERSITY OF SKÖVDE

# *Olika problem vid implementation av filformat och standarder ...*

- Det finns flera olika typer av problem kopplat till **entydighet** och **precision** i tekniska specifikationer av standarder
- Implementationer av en teknisk specifikation av en standard kan **avvika från specifikationen**
- Det finns **influenser mellan** den tekniska **specifikationen** av en standard och dess **implementationer** i programvara
- Bland utvecklare finns **oro för patent** som påverkar möjligheten att implementera standarder i programvara

(Gamalielsson & Lundell, 2013)

# *Sammanfattning – **Öppna frågor** …*
## *Hur kan data och resultat från forskning bevaras?*

- Vet forskare **vilka öppna filformat** som bör användas?
- Används Kammarkollegiets vägledning (som inkluderar ett antal öppna filformat) då forskningsprojekt planeras?
- **Vem analyserar** ett **potentiellt relevant filformat** utifrån ett (tekniskt, juridiskt och ekonomiskt) arkivperspektiv innan formatet används av en forskare? *Har exempelvis forskare och myndigheter införskaffat alla licenser för alla SEPs som belastar de filformat som används?*
- Arkiveras viktiga data och resultat från forskningsprojekt i **originalformat**? *Konvertering av filer mellan olika filformat innan arkivering innebär stor risk för förlust av data.*
- **Vem arkiverar programvaran** som behövs för att kunna tolka och återanvända filer i de filformat som används?


Can you please unlock?

UNIVERSITY OF SKÖVDE

# *Acknowledgements ...*

The LIM-IT project is financially supported by all participating companies, University of Skövde, and the Knowledge Foundation

http://www.his.se/lim-it/partners/

IN PARTNERSHIP WITH THE
Knowledge Foundation

*… acknowledgements extend to several other individuals and partners in a number of (previously conducted) research projects, which have received financial support from several organisations, including: Konkurrensverket, Kammarkollegiet, EU, Vinnova, ...*

UNIVERSITY OF SKÖVDE

# Some sources (1/3) ...


- BSI (2015) What is a Standard?, British Standards, http://www.bsieducation.org/Education/about/what-is-a-standard.shtml
- Cerf, V. G. (2010) Future Imperfect, IEEE Internet Computing, Vol. 14(1), pp. 30-33.
- Fischer, T. and Lundell, B. (2013) Swedish Dissertations: Archived for the Future?, In Proceedings of International Conference on Making Sense of Converging Media (Academic Mindtrek '13), ACM, New York, ISBN 978-1-4503-1992-8, pp. 176-179.
- FLOSSPOLS (2005) Deliverable D4, Open Standards and Interoperability Report: An Economic Basis for Open Standards, flosspols.org
- Gamalielsson, J. & Lundell, B. (2013) Experiences from implementing PDF in open source: challenges and opportunities for standardisation processes, In Jakobs, K. (Ed.) Proc. of the 8th IEEE Conf. on Standardization and Innovation in Information Technology (SIIT 2013), ISBN 3-86130-802-9, IEEE, Piscataway, pp. 39-49.
- Ghosh, P. (2015) Google's Vint Cerf warns of 'digital Dark Age', BBC News, 13 February.
- GOV.UK (2012) Open Standards Principles: For software interoperability, data and document formats in government IT specifications, HM Government, www.gov.uk/government/uploads/system/uploads/attachment_data/file/78892/Open-Standards-Principles-FINAL.pdf
- Jakobs, K. (2000) Standardization processes in IT: Impact, problems and benefits of user participation, Vieweg, Braunschweig, ISBN 978-3-322-86847-3.

UNIVERSITY OF SKÖVDE

# *Some sources (2/3) ...*


- Kammarkollegiet (2015a) Högskolan i Skövde bidrar till Kammarkollegiets riktlinjer för öppna standarder vid upphandling av IT-system, Kammarkollegiet, http://avropa.se/Nyheter/2015/forv/samarbete-med-hogskolan-i-skovde/
- Kammarkollegiet (2015b) Programvaror och tjänsterhandling av IT-system, Kammarkollegiet, https://www.avropa.se/ramavtal/ramavtalsomraden/it-och-telekom/Programvaror-och-tjanster/
- Kammarkollegiet (2016) Open IT-Standards, Kammarkollegiet, 7 mars, Dnr 96-38-2014. http://www.avropa.se/globalassets/open-it-standards.pdf
- Katz, A., Lundell, B. & Gamalielsson, J. (2016) Software, copyright and the learning environment: an analysis of the IT contracts Swedish schools impose on their students and the implications for FOSS, International Free and Open Source Software Law Review, Vol. 8(1), pp. 1-28.
- Konkurrensverket (2016) Nyhetsbrev från Konkurrensverket: Upphandling och konkurrens, 3 november, Swedish Competition Authority. http://www.anpdm.com/newsletter/3907126/43425942794940 5F4071
- Library and Archives Canada (2015) Guidelines on File Formats for Transferring Information Resources of Enduring Value, Library and Archives Canada, 5 February. Canada.
- Lundell, B. (2011) e-Governance in public sector ICT procurement: what is shaping practice in Sweden?, European Journal of ePractice, Vol. 12(6), 66-78.

UNIVERSITY OF SKÖVDE

# *Some sources (3/3) ...*


- Lundell, B. (2012) Why do we need Open Standards?, In Orviska, M. and Jakobs, K. (Eds.) Proceedings 17th EURAS Annual Standardisation Conference 'Standards and Innovation', The EURAS Board Series, Aachen, ISBN: 978-3-86130-337-4, pp. 227-240.
- Lundell, B., Abdurahmanovic, A., Andersson, S., Bergström, E., Feist, J., Gamalielsson, J., Gustavsson, T., Kahlbom, R. & Papaxanthis, K. (2012) How can Open Standards be effectively implemented in Open Source? Challenges and the ORIOS project, In Proc. of the 8th Int. Conf. on Open Source Systems (OSS 2012): IFIP Advances in Information and Communication Technology 378, Springer, ISBN 978-3-642-33441-2, pp. 383-388.
- Lundell, B., Gamalielsson, J. & Katz, A. (2015) On implementation of Open Standards in software: To what extent can ISO standards be implemented in open source software?, International Journal of Standardization Research, Vol. 13(1), pp. 47-73.
- Lundell, B., Gamalielsson, J. & Tengblad, S. (2016) IT-standarder, inlåsning och konkurrens: En analys av policy och praktik inom svensk förvaltning, Uppdragsforskningsrapport 2016:2, Konkurrensverket, ISSN: 1652-8089. www.konkurrensverket.se/nyheter/problem-med-slutna-standarder-vid-it-upphandlingar/
- The National Archives (2011) Suitable file formats for transfer of digital records to The National Archives, September, The National Archives, U.K.
- Smith, D. (2015) Father of the internet: 'If we don't move now, we risk losing all the data we've created in the 21st century', Business Insider UK, 20 February.