Data & Knowledge Management

Mirco Schönfeld
University of Bayreuth

mirco.schoenfeld@uni-bayreuth.de
@TWlyY29

https://commons.wikimedia.org/wiki/File:Floppy_Apple_Macintosh_LC_III_e_Personal_LaserWriter_LW320.jpg
https://commons.wikimedia.org/wiki/File:Floppy_disk_2009_G1.jpg
https://commons.wikimedia.org/wiki/File:MS-DOS_6.22_floppy_disks_20110326.jpg

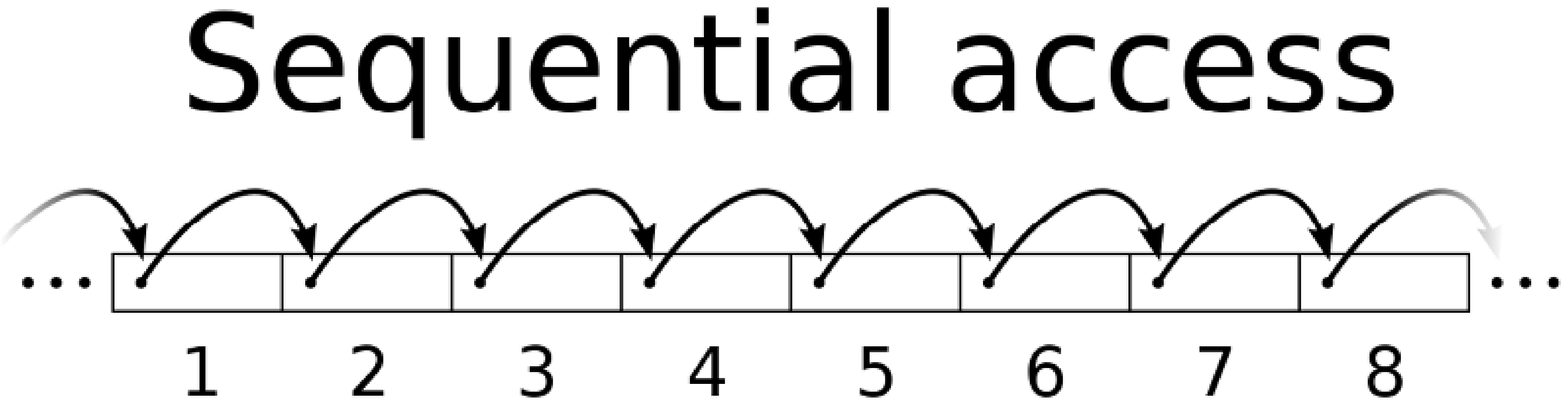
History of Storage Medium

Year	Storage medium	Capacity in Kilobyte	Equivalent in Punch Cards
1890 / 1891	Punch Card	0,08	1
1951	Magnetic tape	800	10.000
1969 - 1975	8 inch floppy disc	80 - 1.000	1.000 - 12.500
1976	5,25 inch floppy disc	110 - 1.200	1.375 - 15.000
1982 - 1998	3,5 inch floppy disc	720 - 2.880	9.000 - 36.000
1982	Compact Disc	650.000 - 900.000	8,125 Mio. - 11,25 Mio.
1994	ZIP-drive	100.000 - 750.000	1.25 Mio. - 9,375 Mio.
1996	USB stick	8.000 - 1.000 Mio.	100.000 - 12.500 Mio.
2001	SD Memory Card	8.000 - 2.000 Mio.	100.000 - 25.000 Mio.
2001	DVD	4,7 Mio. - 18 Mio.	58,75 Mio. - 106,25 Mio.
2006	Blu-Ray	5 Mio. - 50 Mio.	58,75Mio. - 403,8 Mio.

https://wkdiscpress.de/ratgeber/chronik-der-speichermedien/

Sequential Access Memory

Tape drive



Data is stored on tape

(Picture may differ from original product.)

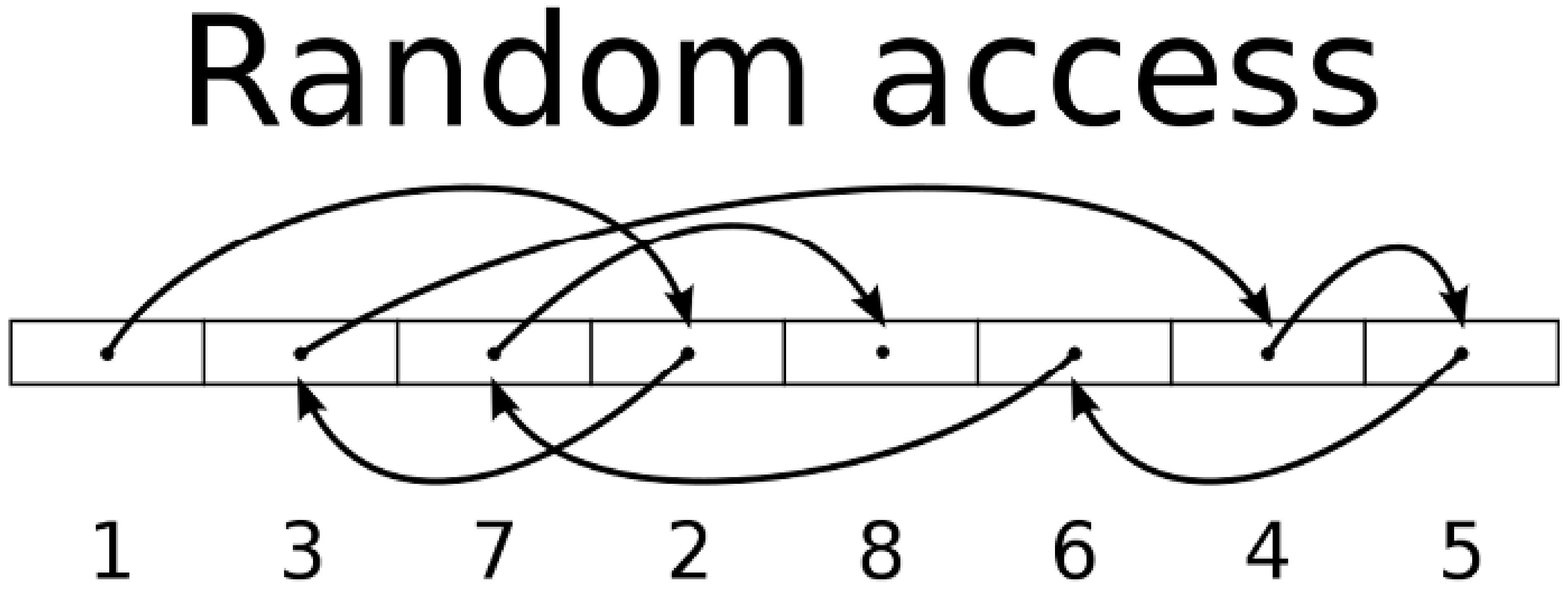
Data is being accessed in a predetermined, ordered sequence.

A data structure is said to have sequential access if one can only visit the values it contains in one particular order.

https://freesvg.org/retro-audio-cassette-and-pencil
https://commons.wikimedia.org/wiki/File:IBM_2401_tape_drives_(1).jpg

Direct Access Storage Device

skip-buttons. awesome.



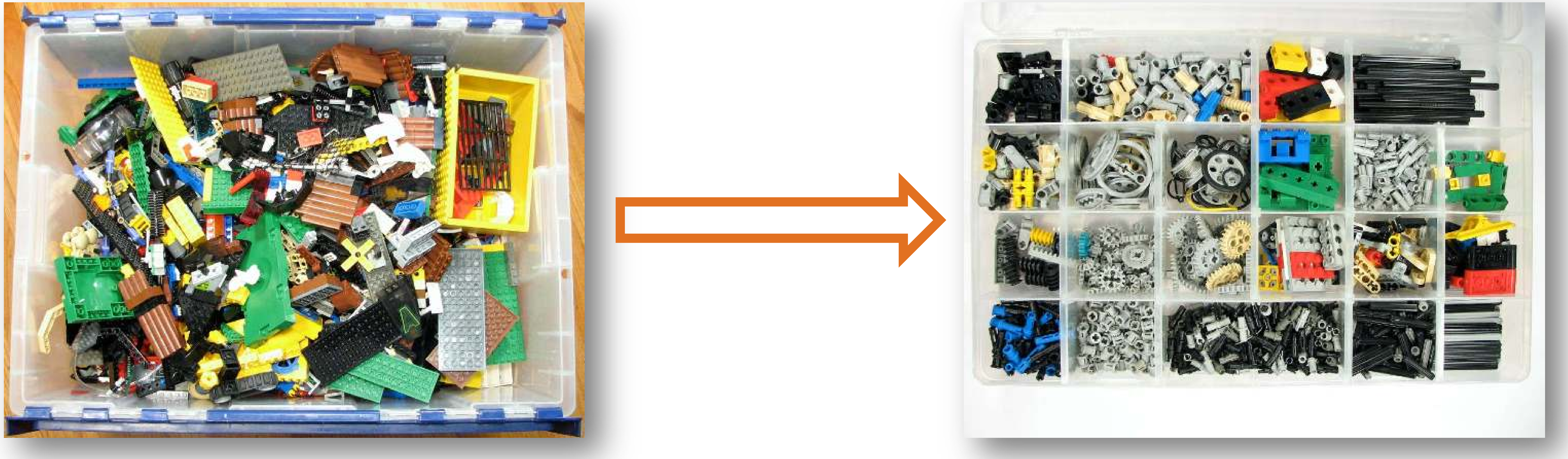
“each physical record has a discrete location and a unique address”

Access an arbitrary element of a sequence in equal time or
any datum from a population of addressable elements roughly as easily and efficiently as any other,
no matter how many elements may be in the set

https://commons.wikimedia.org/wiki/File:Sony-Discman-D-E307CK.jpg
https://commons.wikimedia.org/wiki/File:CD-Scan_20210218.tif

Emergence of Data Management

Blocking: The process of putting data into blocks



This is usually abstracted by a file system (for your hard drive) or a database management system (for your database)

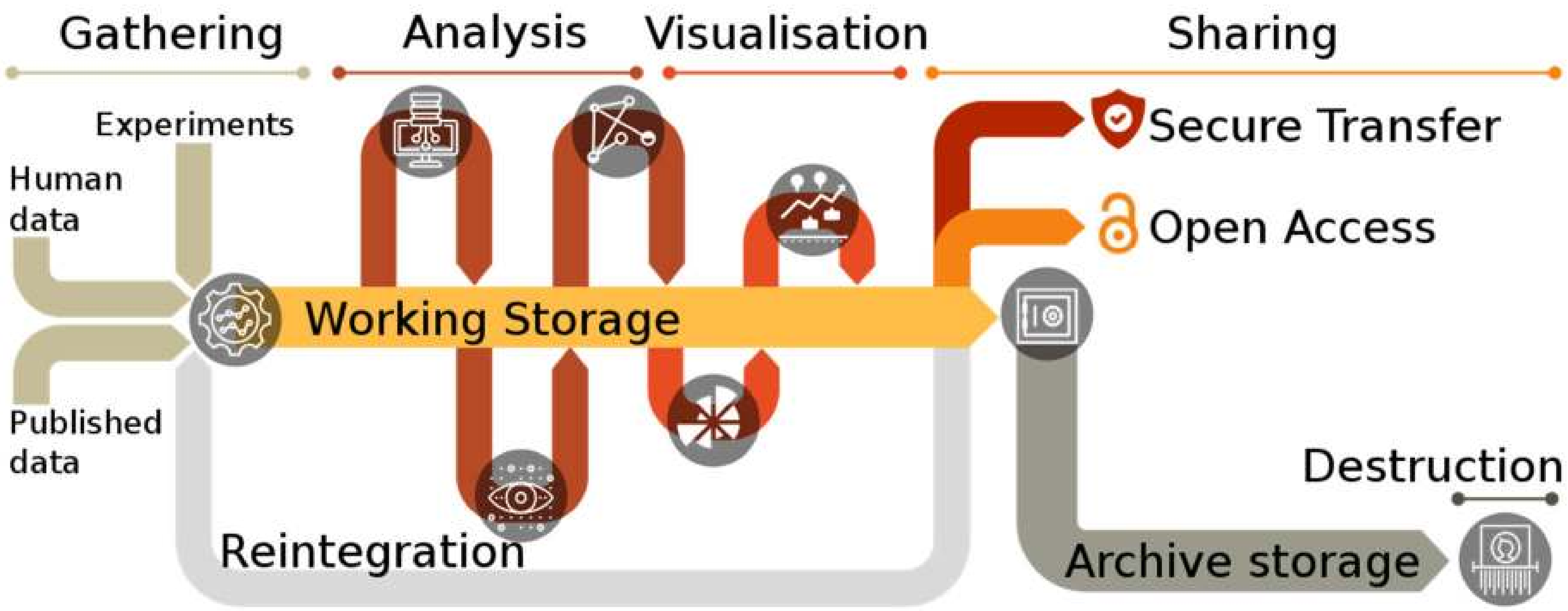
De-blocking: The process of extracting data from blocks

https://www.evilmadscientist.com/2008/how-to-organize-your-lego-bricks-for-efficient-building/

> “Data Management comprises all disciplines related to managing data as a valuable resource

Wikipedia on “Data Management”

Managing Data Lifecycles

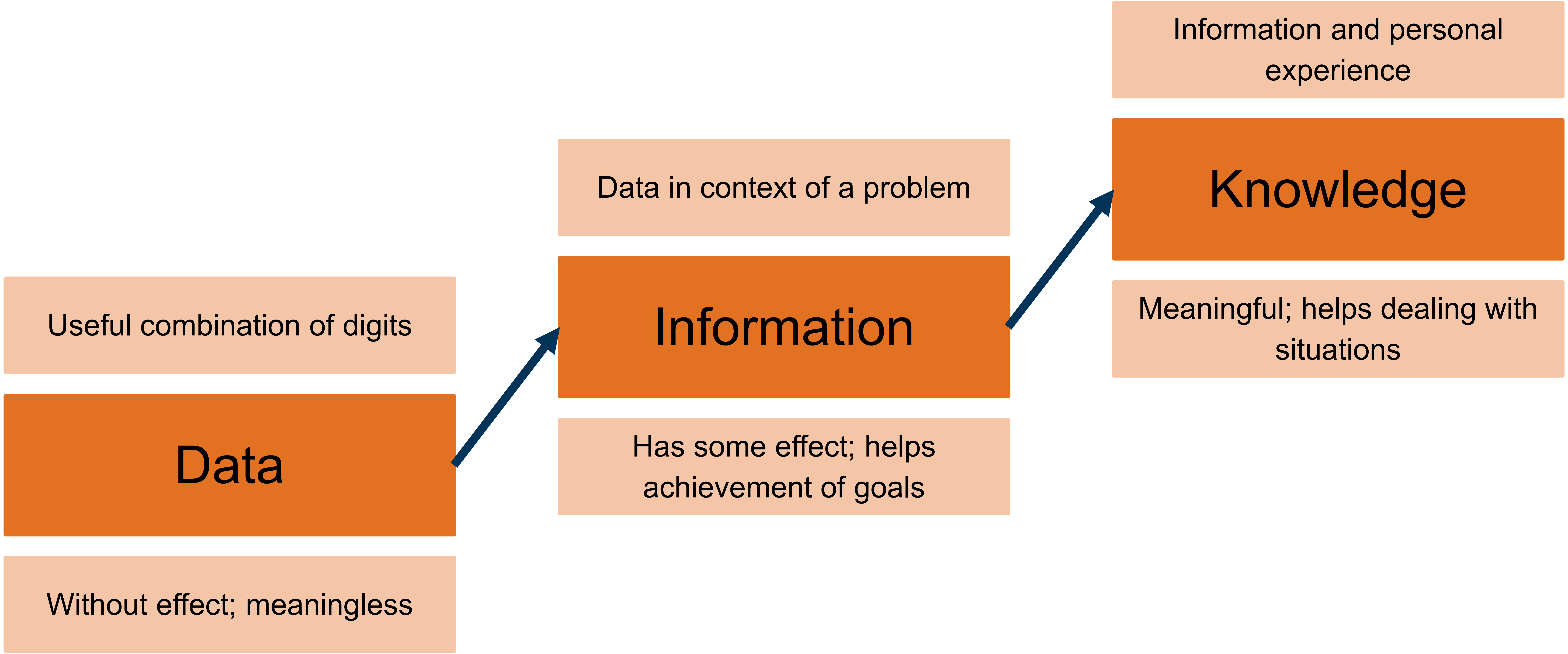


Aspects:

- Data governance
  ensure high data quality
- Data architecture
  models, policies, rules, standards to govern which data is collected, and how it is stored, arranged, integrated, and put to use in data systems and in organizations
- Database and storage management
  administration and organization of databases
- Data security
  govern access and usage of data, protect privacy
- Reference and master data
- Data integration
- Documents and content
- Data warehousing and business intelligence
  strategies and technologies used for analyzing business data; data mining
- Metadata
- Data quality
  does data fit its intended uses in operations, decision making and planning? does it correctly represent the real-world construct to which it refers?

https://commons.wikimedia.org/wiki/File:Data_lifecycle.svg

Data and Knowledge



Useful combination of digits

Data

Without effect; meaningless

Data in context of a problem

Information

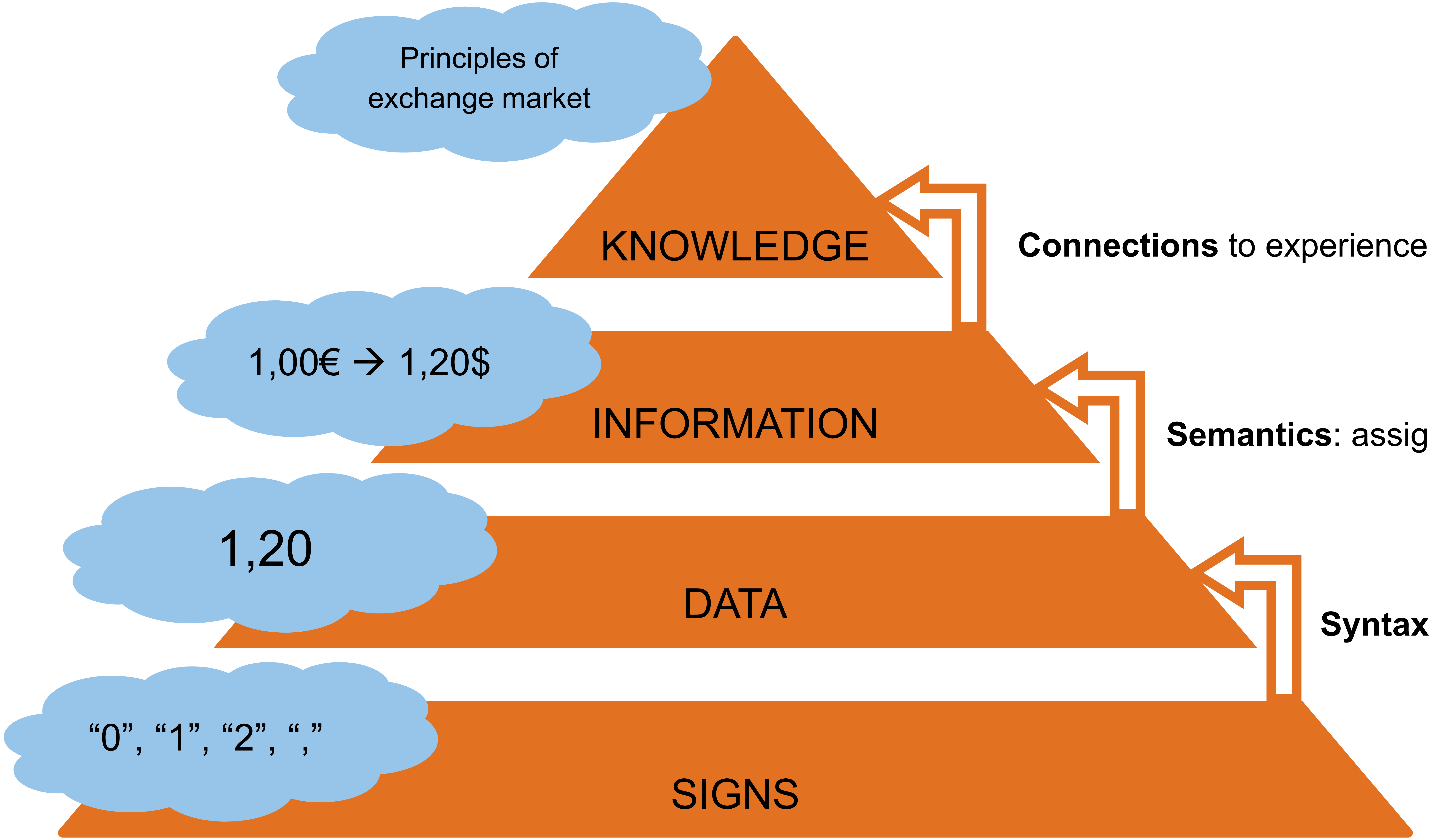
Has some effect; helps achievement of goals

Information and personal experience

Knowledge

Meaningful; helps dealing with situations

From Digits to Knowledge



Connections to experience create knowledge

Semantics: assigning meaning to data

Syntax is used to bring signs into an order

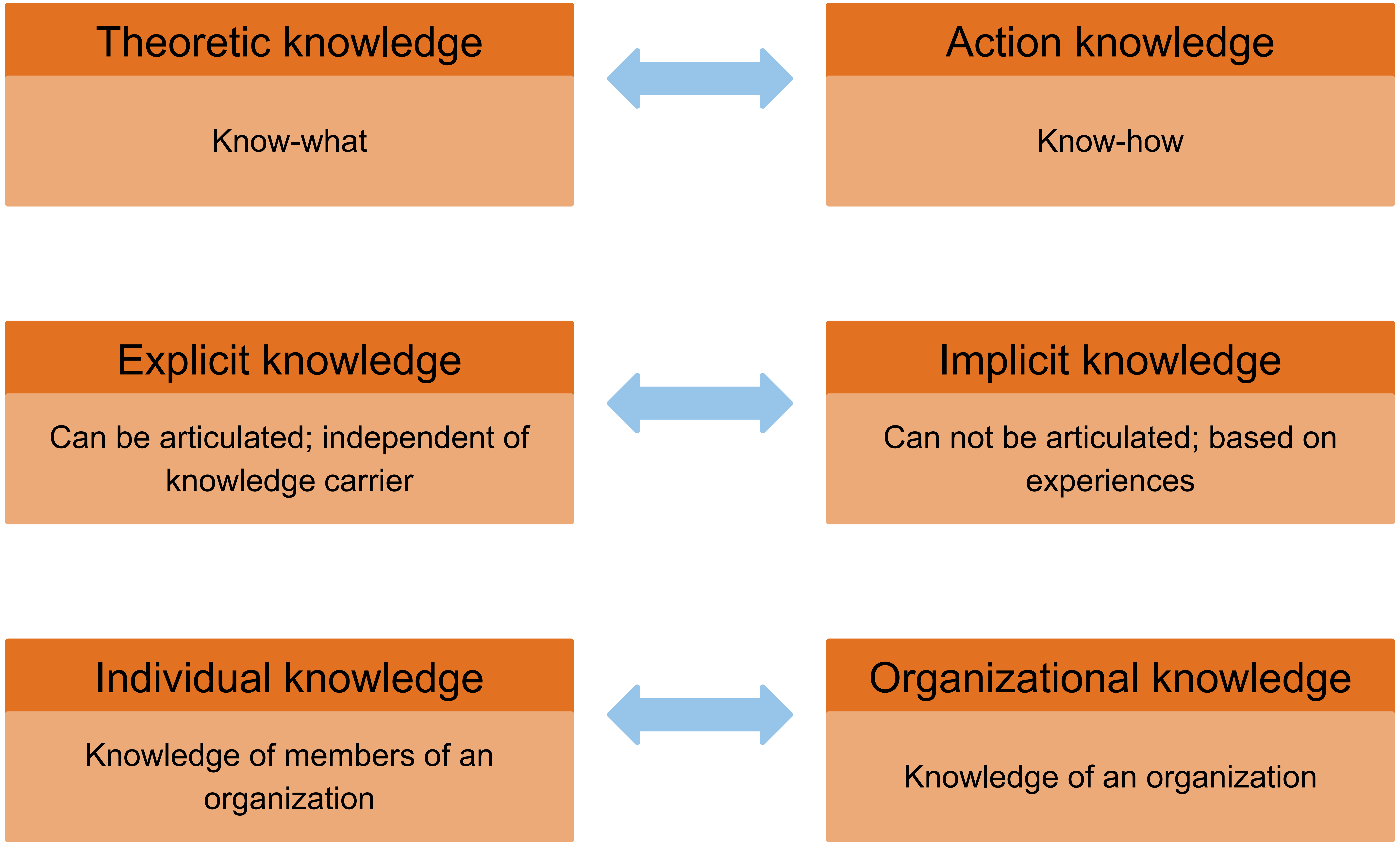
Herrmann, R. (2012). Wissenspyramide. derwirtschaftsinformatiker.de. https://derwirtschaftsinformatiker.de/2012/09/12/it-management/wissenspyramide-wiki/

Knowledge Stairway

Strategic knowledge management
Knowledge-oriented Management
Professional knowledge organisation
Individual Solutions
IT Solutions
Competitiveness
+ uniqueness
Competence
+ act right
Action
+ usage
+ motivation
Knowledge
+ interconnection
Information
+ meaning
Data
+ syntax
Signs
Operational knowledge management

North, K. and Kumta, G., 1998. Wissensorientierte Unternehmensführung. Wiesbaden: Gabler Verlag.

Different Forms of Knowledge



Theoretic knowledge – Know-what ⟷ **Action knowledge** – Know-how

Explicit knowledge – Can be articulated; independent of knowledge carrier ⟷ **Implicit knowledge** – Can not be articulated; based on experiences

Individual knowledge – Knowledge of members of an organization ⟷ **Organizational knowledge** – Knowledge of an organization

Knowledge Management

Knowledge management is the process of creating, sharing, using and managing the knowledge and information of an organization. It refers to a multidisciplinary approach to achieve organizational objectives by making the best use of knowledge.

Knowledge management efforts typically focus on organizational objectives such as improved performance, competitive advantage, innovation, the sharing of lessons learned, integration and continuous improvement of the organization.

Wikipedia on “Knowledge Management”

Core Components of Knowledge Management

Processes / Structure

How to design an organization to facilitate knowledge processes best

People / Culture

How to foster interaction of people and create an environment optimized for knowledge sharing & creation

Technology

How can tools support knowledge sharing and creation

Technological perspective

Technology to support KM

- Groupware
- Content Management Systems
- Workflow Systems
- eLearning
- Project Management Software
- Semantic Technology
- Repositories
- …

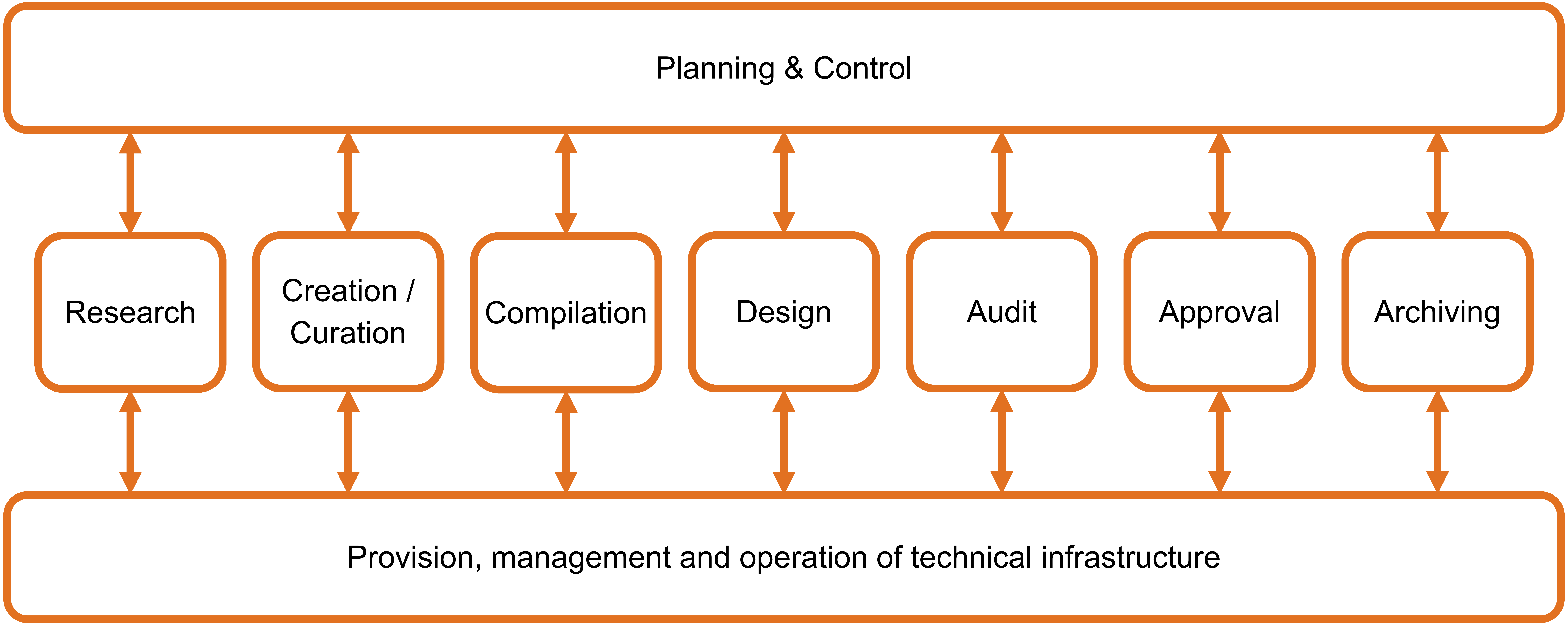
Content Management Systems

Technology and processes to support
collection, management, and publishing of information

Inherently collaborative process consisting of some basic roles and responsibilities

- Creator
  creates and edits content
- Editor
  tuning content message and style of delivery
- Publisher
  releases content for use
- Administrator
  manages access permissions
- Consumer
  views or consumes published content

Functions of Content Management Systems



Planning & Control

Research

Creation / Curation

Compilation

Design

Audit

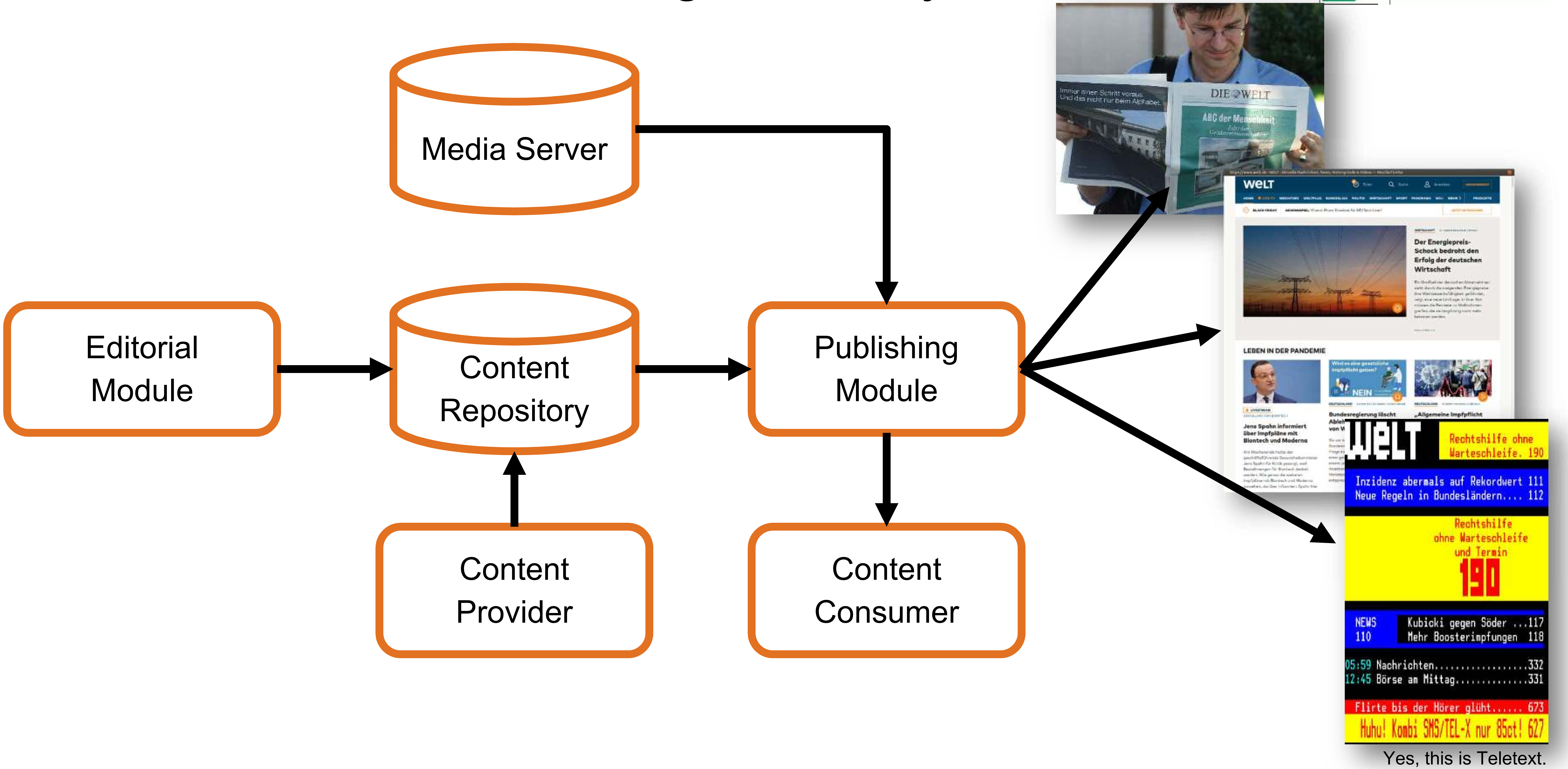
Approval

Archiving

Provision, management and operation of technical infrastructure

Rawolle, J. (2013). Content management integrierter Medienprodukte: ein XML-basierter Ansatz. Springer-Verlag.

Architecture of Content Management Systems



```
WELT  Rechtshilfe ohne
      Warteschleife. 190
Inzidenz abermals auf Rekordwert 111
Neue Regeln in Bundesländern.... 112
        Rechtshilfe
    ohne Warteschleife
        und Termin
           190
NEWS    Kubicki gegen Söder ...117
110     Mehr Boosterimpfungen  118
05:59 Nachrichten.................332
12:45 Börse am Mittag.............331
Flirte bis der Hörer glüht...... 673
Huhu! Kombi SMS/TEL-X nur 85ct! 627
```

Yes, this is Teletext.

https://www.enzyklopaedie-der-wirtschaftsinformatik.de/lexikon/daten-wissen/Informationsmanagement/Informationsmanagement--Konzepte-des/Content-Management-System-/index.html

Version Control Systems

Class of systems responsible for managing changes
to documents or other collections of information

Changes are usually identified by revision levels or “revisions”

Each revision is associated with a timestamp and the person making the change

Revisions can be compared, restored, and, depending on the file type, merged.

Text-based file formats
can be merged. Just saying.

Where to find VCS

Version Control Systems are either standalone or embedded in software

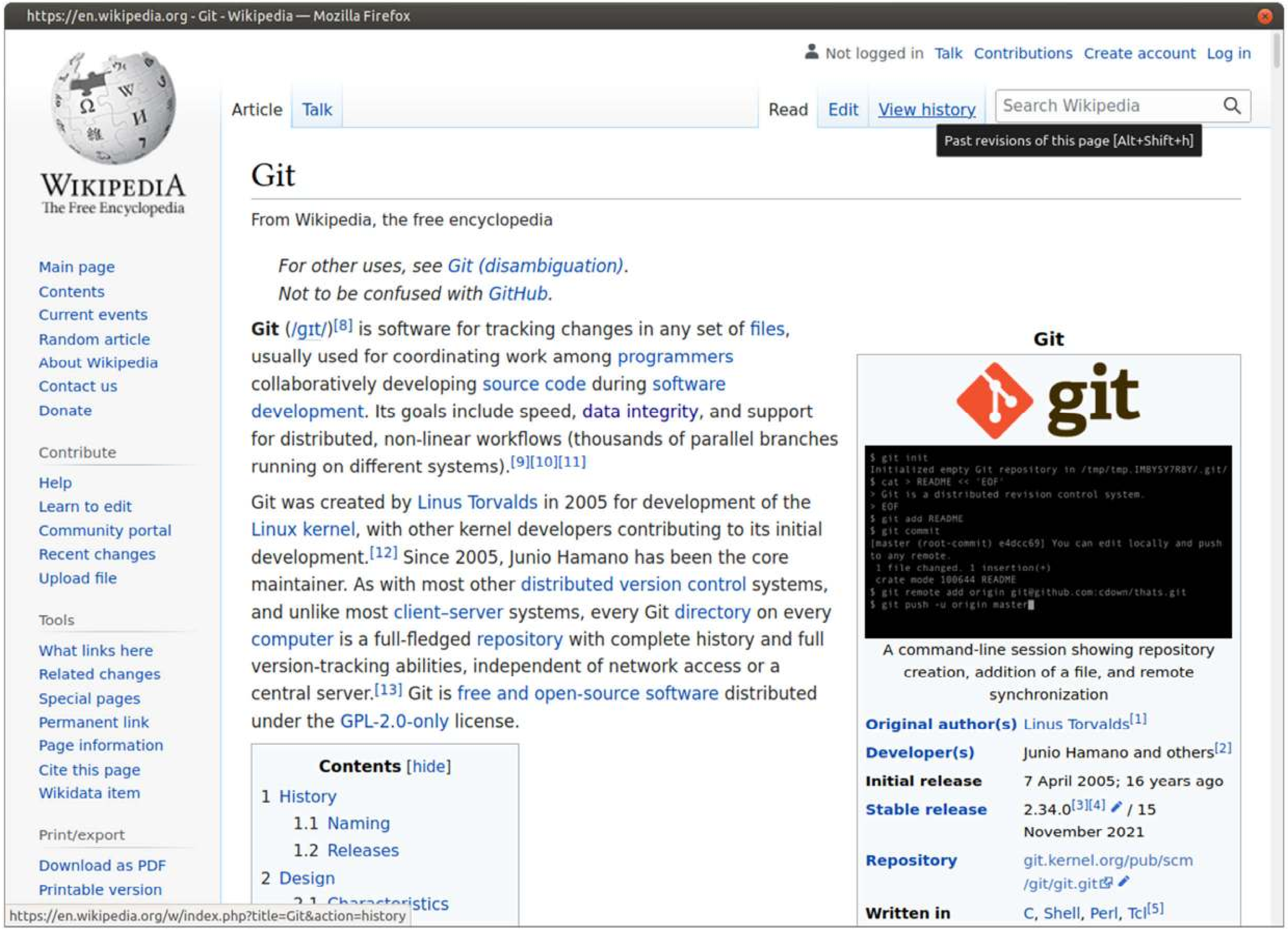
Standalone software:

- Revision Control System (RCS, very old – don’t use)
- Subversion (SVN, old – don’t use)
- Git (use this!)

Software with VCS embedded:

- MediaWiki (Software behind Wikipedia)
- Wordpress (drives ~40% of websites on the internet, they say)
- …

Revisions



Git

From Wikipedia, the free encyclopedia

For other uses, see Git (disambiguation).
Not to be confused with GitHub.

Git (/gɪt/)[8] is software for tracking changes in any set of files, usually used for coordinating work among programmers collaboratively developing source code during software development. Its goals include speed, data integrity, and support for distributed, non-linear workflows (thousands of parallel branches running on different systems).[9][10][11]

Git was created by Linus Torvalds in 2005 for development of the Linux kernel, with other kernel developers contributing to its initial development.[12] Since 2005, Junio Hamano has been the core maintainer. As with most other distributed version control systems, and unlike most client-server systems, every Git directory on every computer is a full-fledged repository with complete history and full version-tracking abilities, independent of network access or a central server.[13] Git is free and open-source software distributed under the GPL-2.0-only license.

Contents [hide]

1 History
  1.1 Naming
  1.2 Releases
2 Design
  2.1 Characteristics

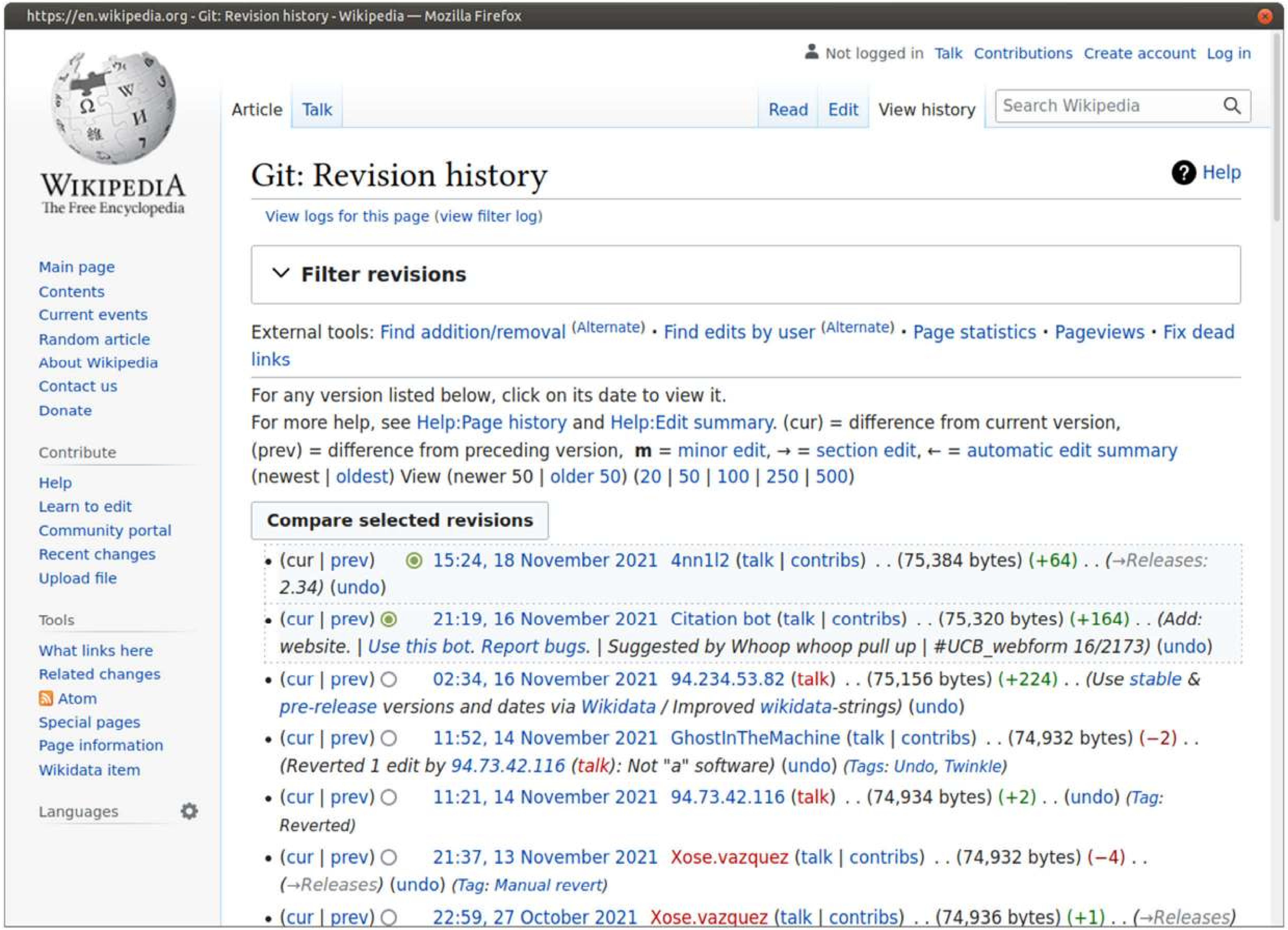
Git

```
$ git init
Initialized empty Git repository in /tmp/tmp.IMBY5Y7R8Y/.git/
$ cat > README << 'EOF'
> Git is a distributed revision control system.
> EOF
$ git add README
$ git commit
[master (root-commit) e4dcc69] You can edit locally and push
to any remote.
 1 file changed, 1 insertion(+)
 crate mode 100644 README
$ git remote add origin git@github.com:cdown/thats.git
$ git push -u origin master
```

A command-line session showing repository creation, addition of a file, and remote synchronization

Original author(s)	Linus Torvalds[1]
Developer(s)	Junio Hamano and others[2]
Initial release	7 April 2005; 16 years ago
Stable release	2.34.0[3][4] / 15 November 2021
Repository	git.kernel.org/pub/scm/git/git.git
Written in	C, Shell, Perl, Tcl[5]

Revisions



Git: Revision history

View logs for this page (view filter log)

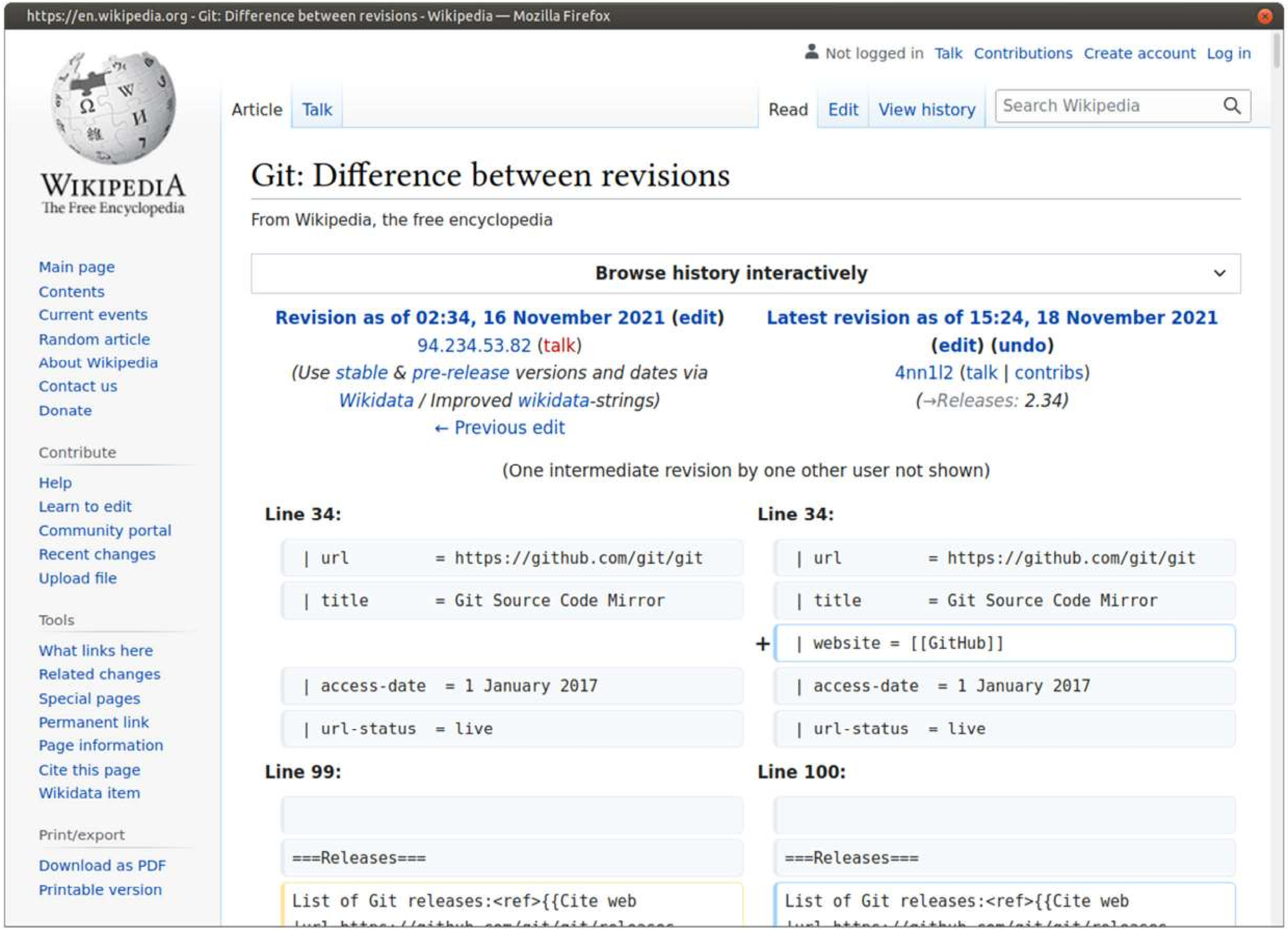
External tools: Find addition/removal (Alternate) • Find edits by user (Alternate) • Page statistics • Pageviews • Fix dead links

For any version listed below, click on its date to view it.
For more help, see Help:Page history and Help:Edit summary. (cur) = difference from current version, (prev) = difference from preceding version, **m** = minor edit, → = section edit, ← = automatic edit summary

(newest | oldest) View (newer 50 | older 50) (20 | 50 | 100 | 250 | 500)

- (cur | prev) ● 15:24, 18 November 2021 4nn1l2 (talk | contribs) . . (75,384 bytes) (+64) . . *(→Releases: 2.34)* (undo)
- (cur | prev) ● 21:19, 16 November 2021 Citation bot (talk | contribs) . . (75,320 bytes) (+164) . . *(Add: website. | Use this bot. Report bugs. | Suggested by Whoop whoop pull up | #UCB_webform 16/2173)* (undo)
- (cur | prev) ○ 02:34, 16 November 2021 94.234.53.82 (talk) . . (75,156 bytes) (+224) . . *(Use stable & pre-release versions and dates via Wikidata / Improved wikidata-strings)* (undo)
- (cur | prev) ○ 11:52, 14 November 2021 GhostInTheMachine (talk | contribs) . . (74,932 bytes) (−2) . . *(Reverted 1 edit by 94.73.42.116 (talk): Not "a" software)* (undo) *(Tags: Undo, Twinkle)*
- (cur | prev) ○ 11:21, 14 November 2021 94.73.42.116 (talk) . . (74,934 bytes) (+2) . . (undo) *(Tag: Reverted)*
- (cur | prev) ○ 21:37, 13 November 2021 Xose.vazquez (talk | contribs) . . (74,932 bytes) (−4) . . *(→Releases)* (undo) *(Tag: Manual revert)*
- (cur | prev) ○ 22:59, 27 October 2021 Xose.vazquez (talk | contribs) . . (74,936 bytes) (+1) . . *(→Releases)*

Revisions



Git: Difference between revisions

From Wikipedia, the free encyclopedia

Browse history interactively

Revision as of 02:34, 16 November 2021 (edit) 94.234.53.82 (talk) *(Use stable & pre-release versions and dates via Wikidata / Improved wikidata-strings)* ← Previous edit	**Latest revision as of 15:24, 18 November 2021 (edit) (undo)** 4nn1l2 (talk | contribs) *(→Releases: 2.34)*

(One intermediate revision by one other user not shown)

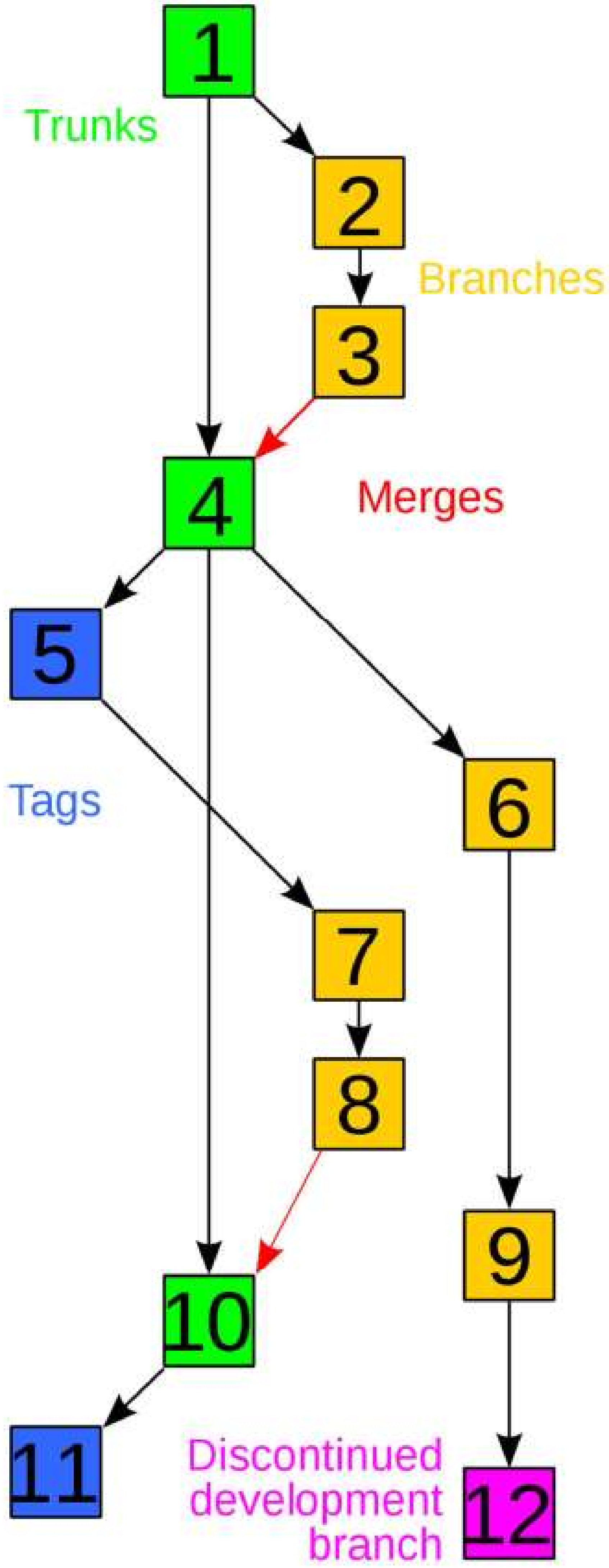
Line 34:		**Line 34:**
`| url          = https://github.com/git/git`		`| url          = https://github.com/git/git`
`| title        = Git Source Code Mirror`		`| title        = Git Source Code Mirror`
	+	`| website = [[GitHub]]`
`| access-date   = 1 January 2017`		`| access-date   = 1 January 2017`
`| url-status   = live`		`| url-status   = live`
Line 99:		**Line 100:**
`===Releases===`		`===Releases===`
`List of Git releases:<ref>{{Cite web`		`List of Git releases:<ref>{{Cite web`

Trunks and Branches

Trunk
Unnamed branch of a file tree under revision control

Structure
The structure of the revisions is not a tree (although it is often referred to as the revision tree) but a directed acyclic graph.

Tag
A tag assigns a label to a revision (including many files) allowing to directly jump to that revision. Often used to label a specific version of a software.



Branching
Duplication of an object under version control. Objects can then be modified separately and in parallel so that they become different. These objects are called branches.

Merge
A fundamental operation that reconciles multiple changes made to a version-controlled collection of files. Necessary when files are modified on two independent branches. The result is a single collection of files that contains both sets of changes.

https://commons.wikimedia.org/wiki/File:Revision_controlled_project_visualization.svg

Long Term Archiving

Long Term Archiving of Data

UNIVERSITÄT BAYREUTH

> For digital preservation, "long term" does not mean issuing a guarantee for five or fifty years, but rather the responsible development of strategies that can cope with the constant changes caused by the information market.
>
> The meaning of "archiving" is more than just the permanent storage of digital information on a data carrier. Rather, it includes the preservation of the permanent availability and thus the subsequent use and interpretability of digital resources.

Heike Neuroth in Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Nestor, 2010.
http://www.nestor.sub.uni-goettingen.de/handbuch/

Goals of long-term archiving

- Long-term, secure storage of the data
- Preserving the interpretability of the data
- Ensure discoverability of data
- Ensure traceability of data

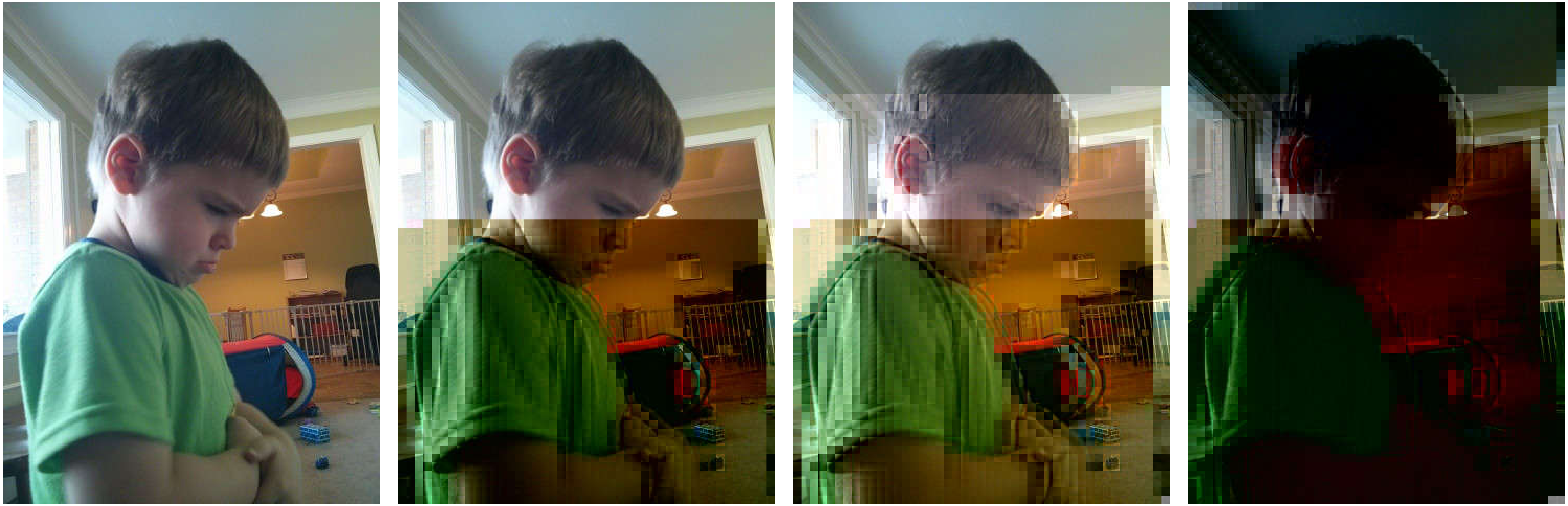
Long-term archiving is more than a backup!

Weber, A. & Piesche, C. (2021). Datenspeicherung, -kuration und Langzeitverfügbarkeit. In M. Putnings, H. Neuroth & J. Neumann (Ed.), Praxishandbuch Forschungsdatenmanagement (pp. 327-356).

Important aspects of long-term archiving

1. Archival: preservation of data substance often called bit-stream preservation.

Illustration of Bit Rot: 4 versions of the same image file consisting of 326272 bits.



original

1 / 326272 bits flipped

2 / 326272 bits flipped

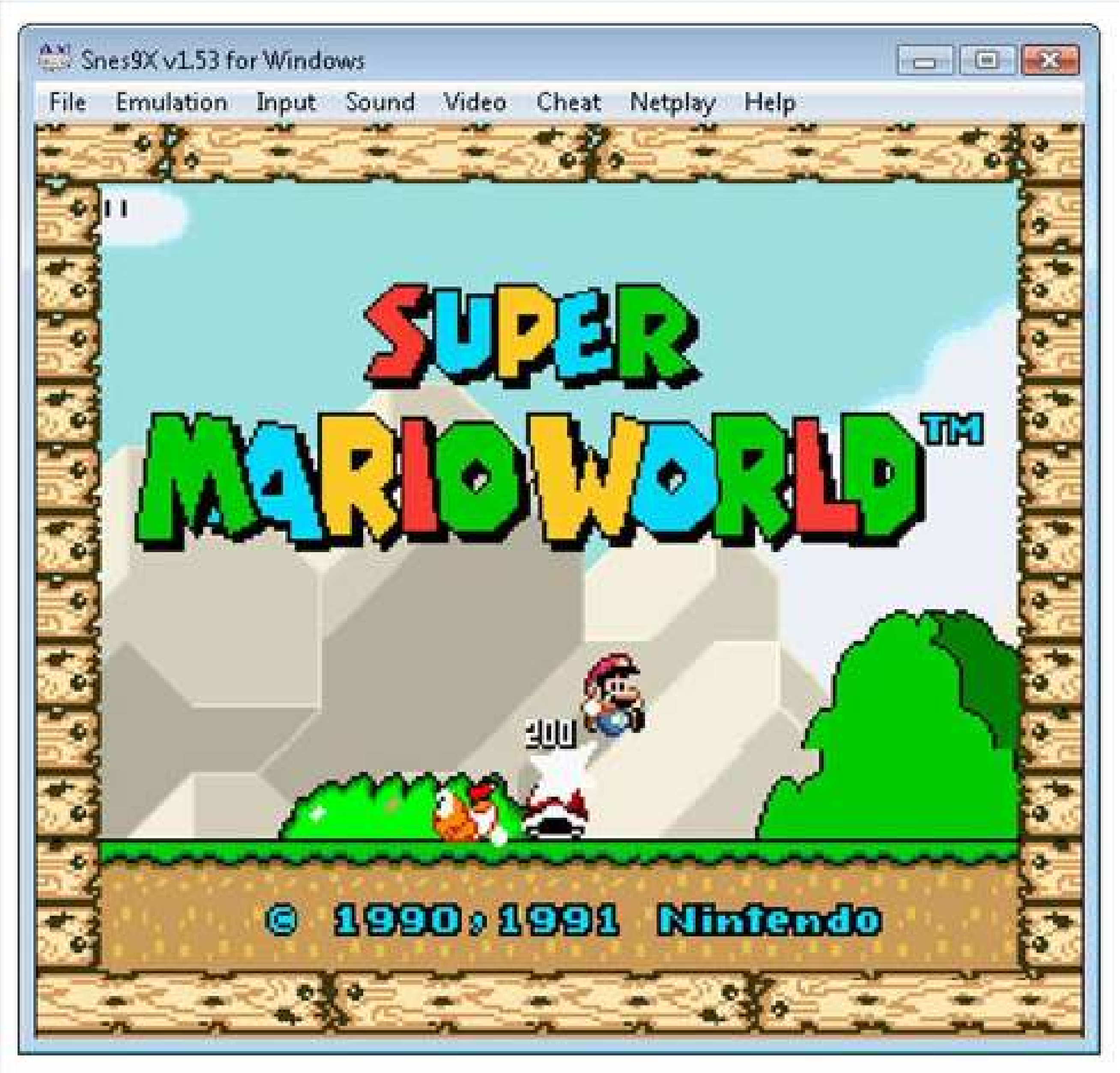
3 / 326272 bits flipped

https://en.wikipedia.org/wiki/Data_degradation

Important aspects of long-term archiving

2. Reusability: preservation of usability
   - Usage of standards; require documentation
   - Migration to current file formats (and open standards!)
   - Preservation of creation context (e.g. software and hardware)

Either keep all the required hard- and software . . . . or emulate the context artificially.



https://www.reddit.com/r/snes/comments/ifuwn8/how_you_like_my_setup/
https://snes9x.de.malavida.com
https://commons.wikimedia.org/wiki/File:SNES_800.jpg

Suitable File Formats

Document type	Format name	File extension
Audio	Waveform Audio	*.wav
	MPEG 1/2 Audio Layer 3	*.mp3
Video	Motion JPEG 2000	*.mj2, *.mjp2
	Matroska Multimedia Container (FF video codec 1)	*.mkv
Images / Raster Graphics	Tagged Image File Format	*.tiff
	Windows Bitmap	*.bmp
	Portable Network Graphics	*.png
Portable Document Format	Acrobat PDF/A - Portable Document Format 1a – 2u	*.pdf
Independent text-based format	Character-Separated Values	*.csv / *.tsv
	Markdown	*.md
	Text File	*.txt
	Extensible Markup Language	*.xml
Office files	None	

https://www.hbz-nrw.de/produkte/langzeitverfuegbarkeit/langzeitverfuegbarkeit-fuer-hochschulen/lzv-dateiformatliste

Thanks.

mirco.schoenfeld@uni-bayreuth.de

Knowledge Representation

Semantic Networks

Important class of representation of knowledge

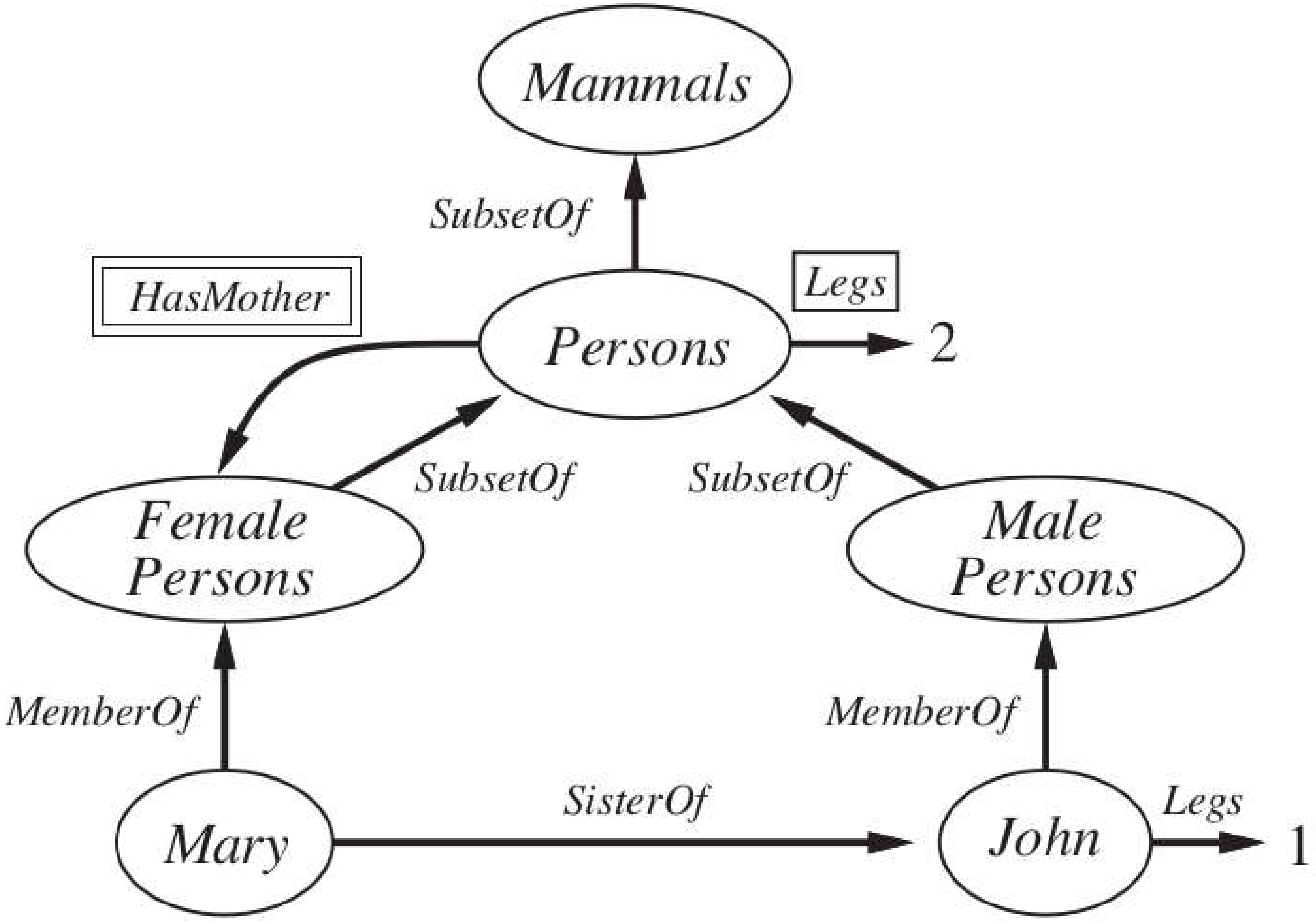
Origin: Charles Peirce “Existential Graphs”

Characteristics:

- Nodes represent concepts
- Nodes are labeled

  Labels specify concepts
- Links specify relations

  is-a, has-a, property-of
- Links are directed
- Inheritance



Peirce, C. S. ,1909. *Existential graphs*. Unpublished manuscript; reprinted in (Buchler 1955).
Markman, A.B., 2013. *Knowledge representation*. Psychology Press.
Russell, S. and Norvig, P., 2002. *Artificial intelligence: a modern approach*. New Jersey: Pearson Education.