HDF5: The most versatile container for sharing scientific and engineering data

Gerd Heber, The HDF Group

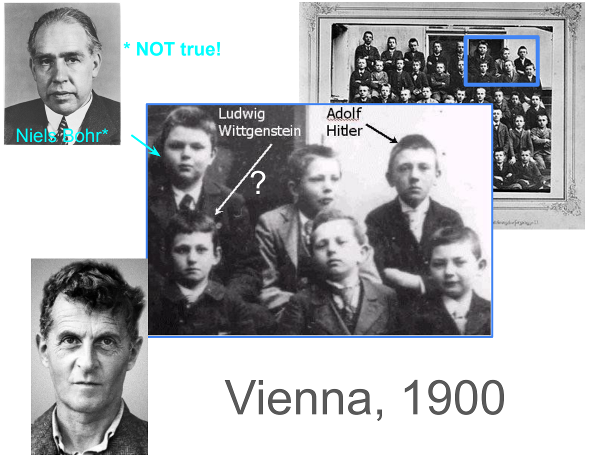
Key Points

A big **thank you** to DESY, the Organizing Committee, our sponsors, and countless helpers

- HDF5 is well-positioned as a versatile and evolving container for structured scientific and engineering data
- Like street lights, HDF5 is a public good (non-rivalrous, non-excludable)
- It’s evolution is intertwined with another public good: language
- These are exciting times for anyone interested in either or both!
- Public goods face unique challenges: Let’s work them together!

Facts inform. Logic convinces. But stories move us. So let me tell you a story.

What could possibly go wrong?



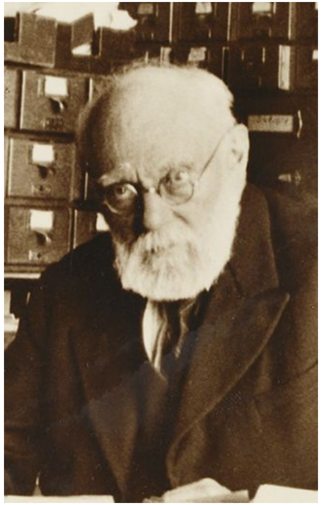
Vienna, 1900

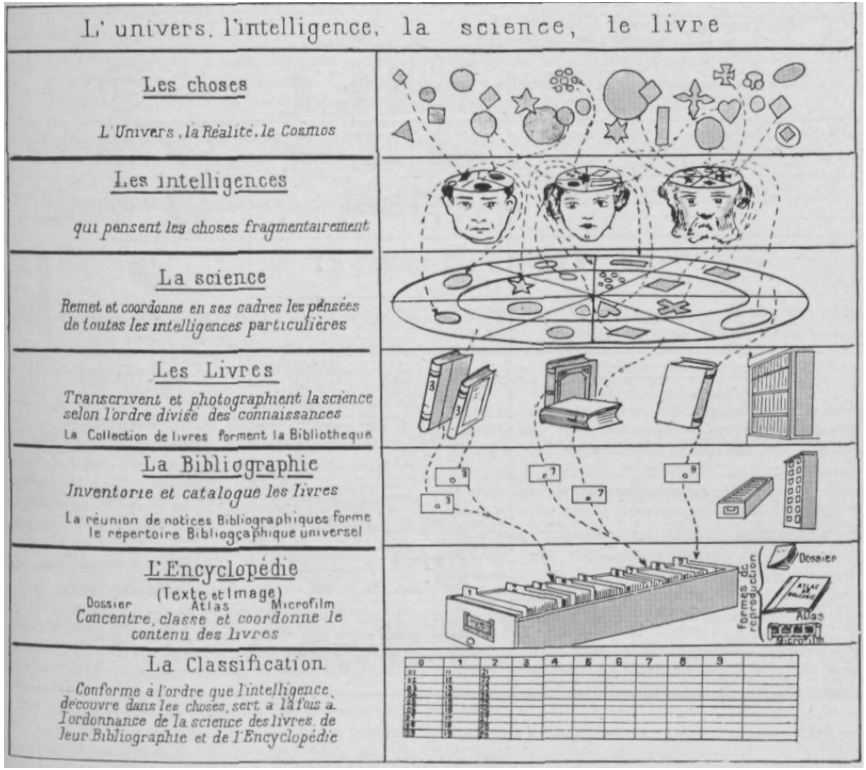
"license with dates and details ... has been, though not unsparingly, indulged" (Edward Bulwer-Lytton)

Paul Otlet

Traité de documentation: le livre sur le livre, théorie et pratique

Brussels, 1934





L' univers. l'intelligence, la science, le livre

Les choses

L'Univers . la Réalité. le Cosmos

Les intelligences

qui pensent les choses fragmentairement

La science

Remet et coordonne en ses cadres les pensées de toutes les intelligences particulières

Les Livres

Transcrivent et photographient la science selon l'ordre divisé des connaissances

La Collection de livres forment la Bibliotheque

La Bibliographie

Inventorie et catalogue les livres

La réunion de notices Bibliographiques forme le repertoire Bibliogçaphique universel

L'Encyclopédie

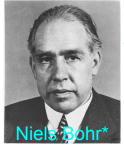
(Texte et Image)

Dossier Atlas Microfilm

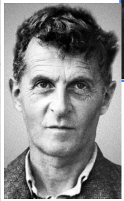
Concentre, classe et coordonne le contenu des livres

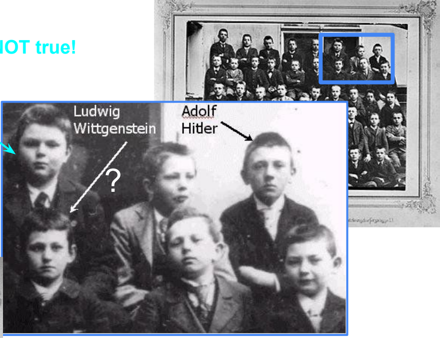
La Classification

Conforme à l'ordre que l'intelligence découvre dans les choses, sert à la fois à l'ordonnance de la science des livres. de leur Bibliographie et de l'Encyclopédie



* NOT true!

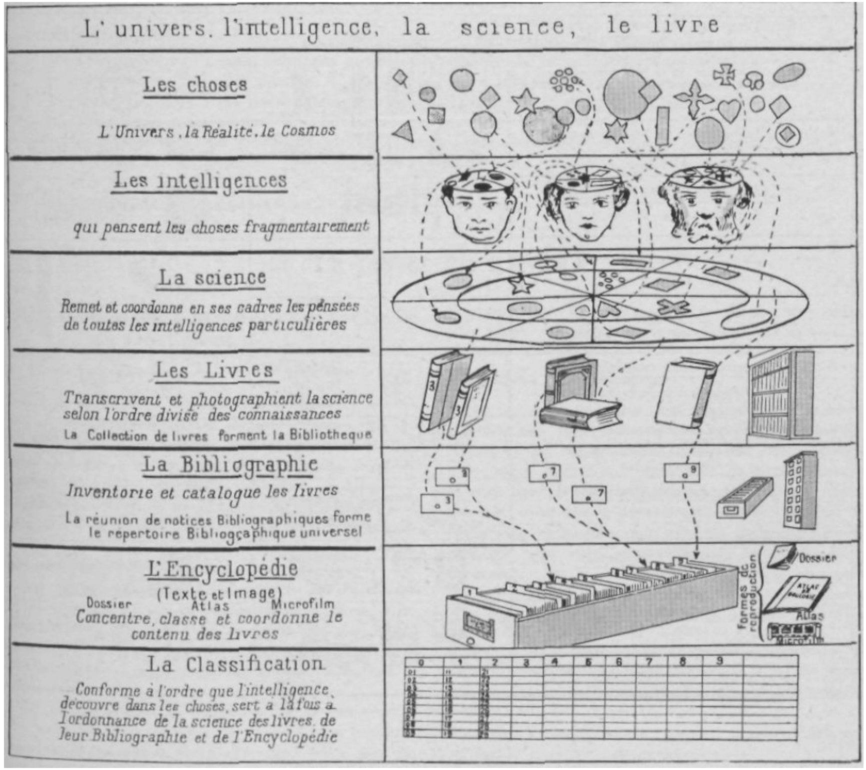




Vienna, 1900

Flakturm IV, Heiligengeistfeld, Hamburg, 1942
(~ 8 km southeast of DESY)

Trinity, 1945
Ivy Mike, 1952



L' univers, l'intelligence, la science, le livre

Les choses

L'Univers, la Réalité, le Cosmos

Les intelligences

qui pensent les choses fragmentairement

La science

Remet et coordonne en ses cadres les pensées de toutes les intelligences particulières

Les Livres

Transcrivent et photographient la science selon l'ordre divisé des connaissances

La Collection de livres forment la Bibliotheque

La Bibliographie

Inventorie et catalogue les livres

La réunion de notices Bibliographiques forme le repertoire Bibliographique universel

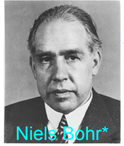
L'Encyclopédie

(Texte et Image)
Dossier Atlas Microfilm

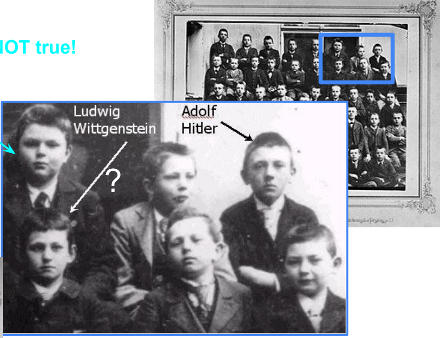
Concentre, classe et coordonne le contenu des livres

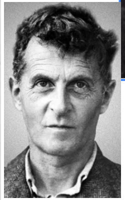
La Classification

Conforme à l'ordre que l'intelligence découvre dans les choses, sert à la fois à l'ordonnance de la science des livres, de leur Bibliographie et de l'Encyclopédie



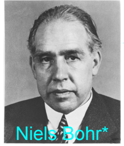
* NOT true!



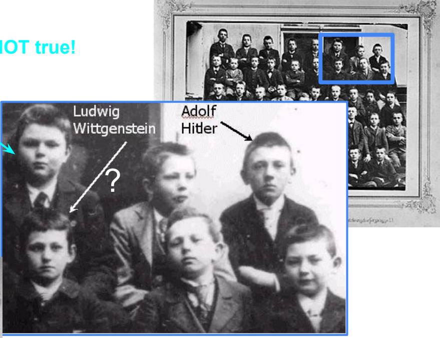


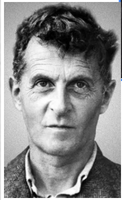
Vienna, 1900

HDF5 1.0.0 November 6, 1998

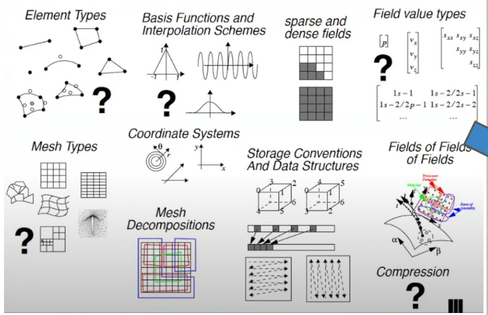


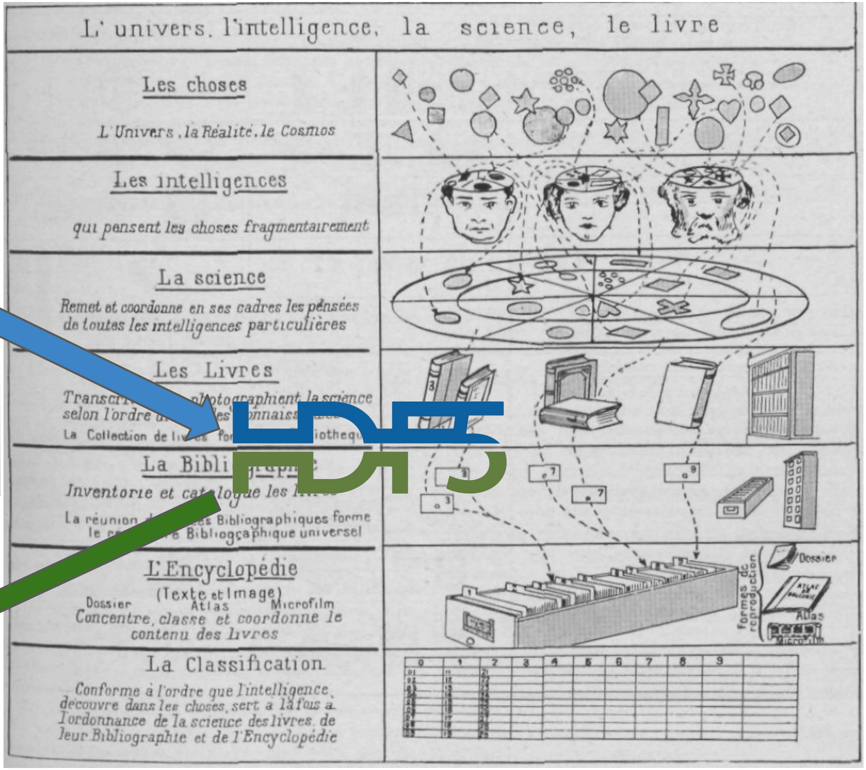
* NOT true!





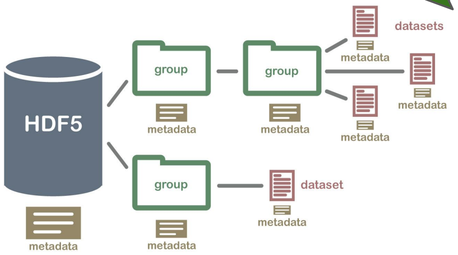
Vienna, 1900





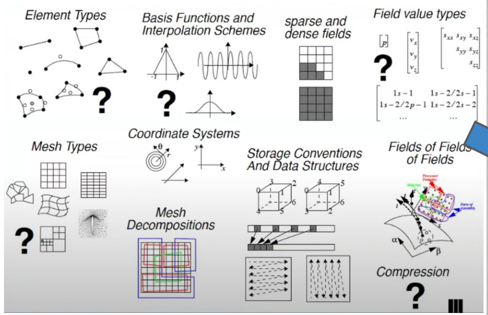
L' univers. l'intelligence, la science, le livre

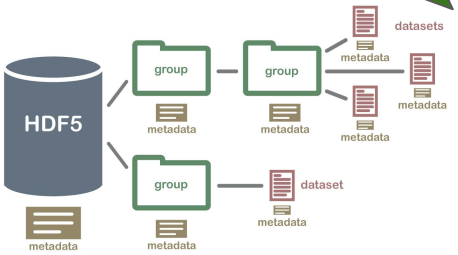
Les choses L'Univers. la Réalité. le Cosmos	
Les intelligences qui pensent les choses fragmentairement	
La science Remet et coordonne en ses cadres les pensées de toutes les intelligences particulières	
Les Livres Transcri... photographient la science selon l'ordre ... des ...onnais... La Collection de livres for... ...iothèq...	
La Bibli... Inventorie et catalogue les livres La réunion ...s Bibliographiques forme le ... Bibliographique universel	
L'Encyclopédie (Texte et Image) Dossier Atlas Microfilm Concentre, classe et coordonne le contenu des livres	
La Classification Conforme à l'ordre que l'intelligence découvre dans les choses, sert à la fois à l'ordonnance de la science des livres, de leur Bibliographie et de l'Encyclopédie	

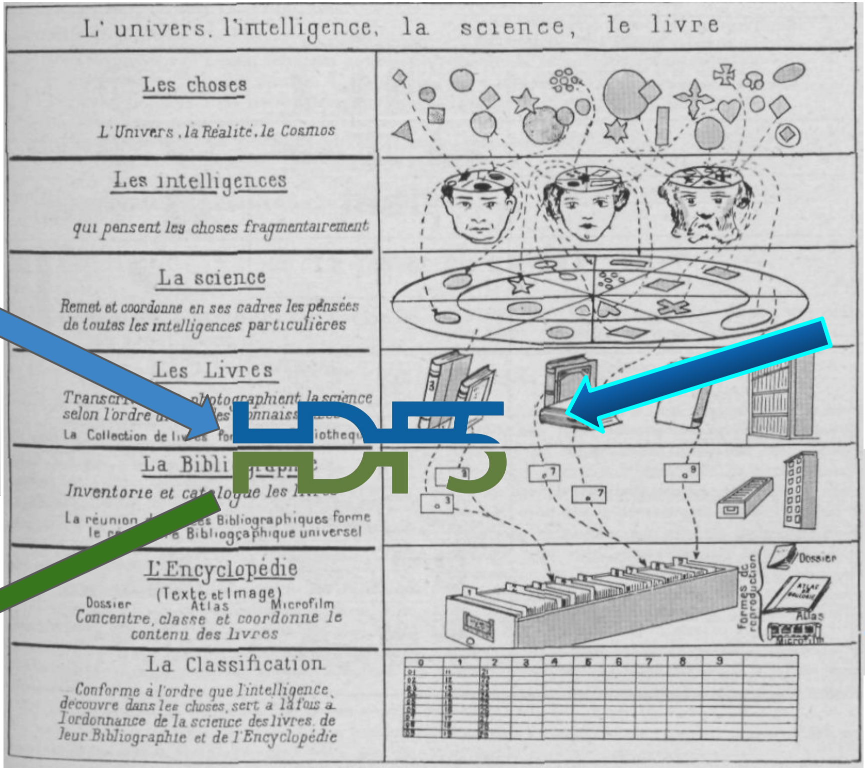


https://www.youtube.com/@hdf5

Wittgenstein's Turn and Quantum 2nd







L' univers. l'intelligence, la science, le livre

Les choses

L'Univers. la Réalité. le Cosmos

Les intelligences

qui pensent les choses fragmentairement

La science

Remet et coordonne en ses cadres les pensées de toutes les intelligences particulières

Les Livres

Transcri[illegible] photographient la science selon l'ordre [illegible] les connaiss[illegible]

La Collection de livres for[illegible]iothequ[illegible]

La Bibli[illegible]

Inventorie et catalogue les li[illegible]

La réunion d[illegible]es Bibliographiques forme le r[illegible]e Bibliographique universel

L'Encyclopédie

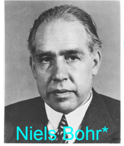
(Texte et Image)

Dossier Atlas Microfilm

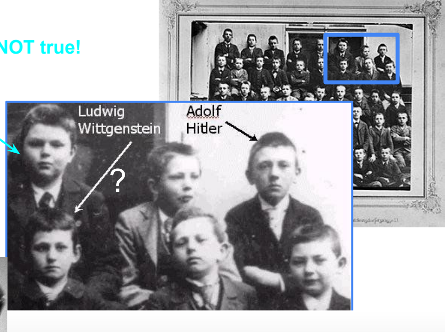
Concentre, classe et coordonne le contenu des livres

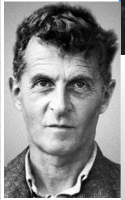
La Classification

Conforme à l'ordre que l'intelligence découvre dans les choses, sert à la fois à l'ordonnance de la science des livres, de leur Bibliographie et de l'Encyclopédie



* NOT true!

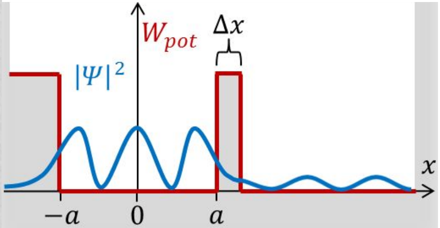




$P(w_n|w_1, w_2, \ldots, w_{n-1})$ *

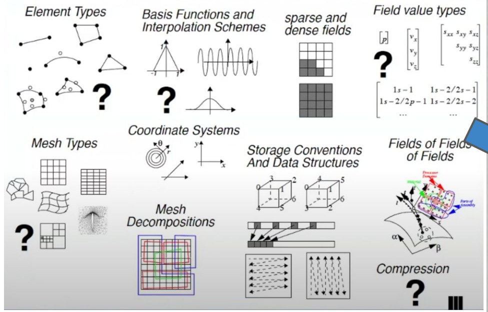
"People's language reflects their reality;
LLM's language reflects people's language."

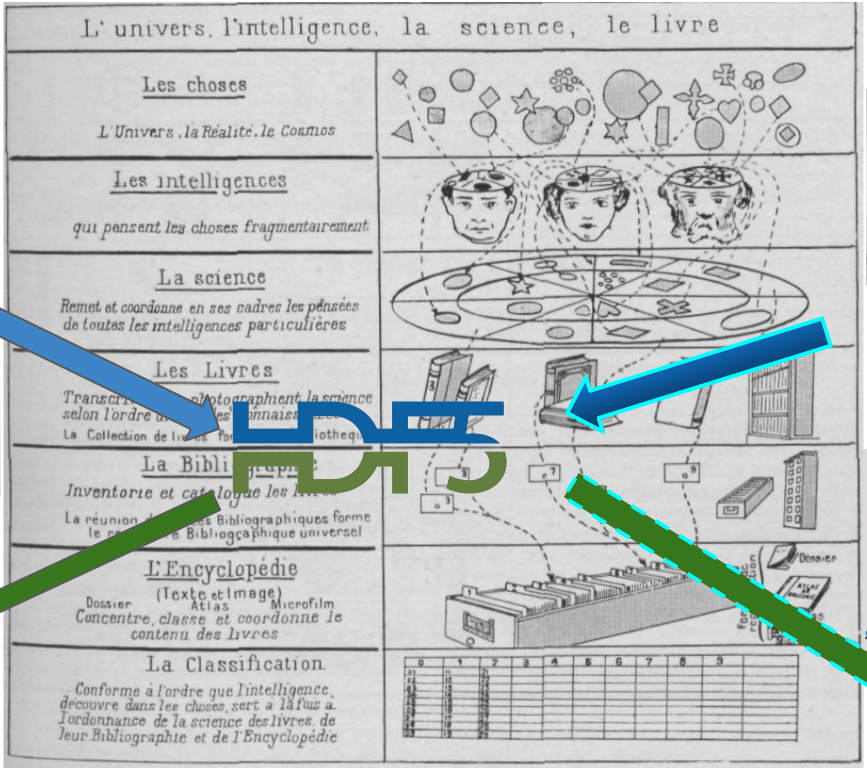
* What's next?

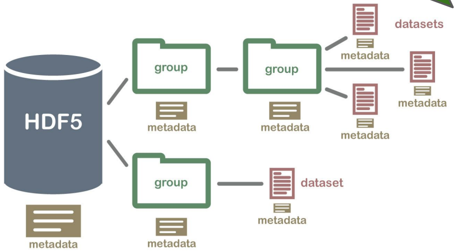


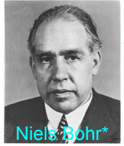
"It is the stillest voice that brings on the storm. Thoughts that come on doves' feet guide the world." (Friedrich Nietzsche)

X-ready HDF5?

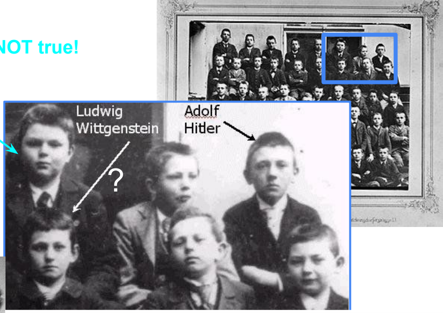


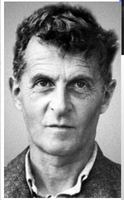

L' univers. l'intelligence, la science, le livre
Les choses
L'Univers. la Réalité. le Cosmos
Les intelligences
qui pensent les choses fragmentairement
La science
Remet et coordonne en ses cadres les pensées de toutes les intelligences particulières
Les Livres
Transcri… photographient la science selon l'ordre … des connaiss…
La Collection de livres for… …iothequ…
La Bibli…
Inventorie et catalogue les …
La réunion … Bibliographiques forme le … Bibliographique universel
L'Encyclopédie
(Texte et Image)
Dossier
Atlas
Microfilm
Concentre, classe et coordonne le contenu des livres
La Classification
Conforme à l'ordre que l'intelligence découvre dans les choses, sert à la fois à l'ordonnance de la science des livres, de leur Bibliographie et de l'Encyclopédie
Dossier
0
1
2
3
4
5
6
7
8
9
HDF5





* NOT true!

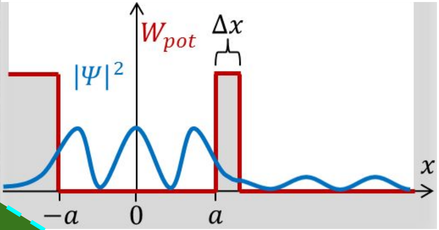




$P(w_n|w_1, w_2, \ldots, w_{n-1})$ *

"People's language reflects their reality;
LLM's language reflects people's language."

* What's next?



?

Cloud-ready HDF5

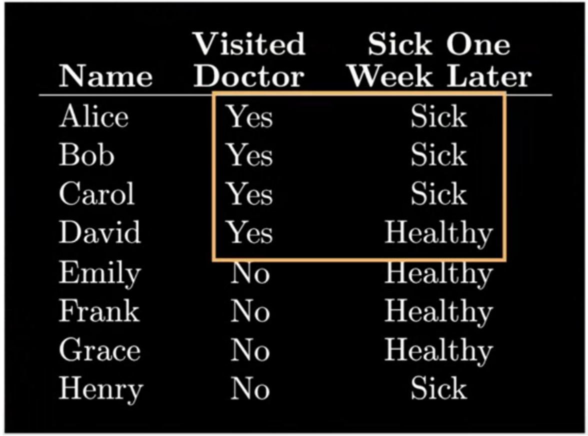
Cloud-ready HDF5 encompasses two complementary approaches for efficiently working with HDF5 data in cloud environments:

- Cloud-Optimized HDF5 (COH5) is a practice that restructures HDF5 files to enable efficient partial reading from S3-compatible object stores.
- HDF5 Scalable Data Service (HSDS) is a REST-based service that provides HDF5 functionality through HTTP APIs.
  - Stores HDF5 data in an object storage-friendly format
  - Multi-reader/multi-writer (MRMW) support
  - Supports horizontal scaling across multiple service instances
  - Offers fine-grained access control and authentication

COH5 focuses on how the data is structured at rest, HSDS provides the mechanism for dynamic access and scaling in a cloud environment. HSDS can leverage COH5 files for even better performance, but it can also work with non-optimized HDF5 files by intelligently managing their access from object storage.

AI/ML-ready HDF5

Does going to the doctor make you sick?



Name	Visited Doctor	Sick One Week Later
Alice	Yes	Sick
Bob	Yes	Sick
Carol	Yes	Sick
David	Yes	Healthy
Emily	No	Healthy
Frank	No	Healthy
Grace	No	Healthy
Henry	No	Sick

AI-ready HDF5 refers to HDF5 containers that are structured using HDF5's hierarchical model and strategically designed, annotated, and managed to meet the specific requirements of AI/ML workflows. This includes optimizing for efficient data access and processing by AI algorithms, ensuring data quality and integrity, providing comprehensive and standardized metadata for context and reproducibility, facilitating seamless integration with AI tools and frameworks, and adhering to FAIR data principles to maximize utility and collaboration.

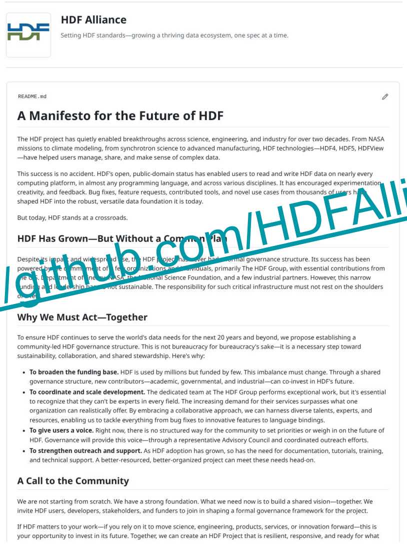
Community-ready HDF5

An evolution of the HDF5 ecosystem that is structured, governed, and developed in a way that actively invites, enables, and values contributions from a broad and diverse community of users, developers, and stakeholders. It goes beyond open source to embody principles of openness, collaboration, transparency, and sustainability. A community-ready HDF5 would demonstrate the following characteristics:

1. Transparent and Inclusive Governance
2. Contribution-Friendly Infrastructure
3. Open and Responsive Development Process
4. Roadmap Shaped by Community Needs
5. Support for Ecosystem Collaboration
6. Accessible Education and Outreach
7. Trustworthy and Accountable

This is THE challenge!

https://github.com/HDFAlliance



HDF Alliance

Setting HDF standards—growing a thriving data ecosystem, one spec at a time.

README.md

A Manifesto for the Future of HDF

The HDF project has quietly enabled breakthroughs across science, engineering, and industry for over two decades. From NASA missions to climate modeling, from synchrotron science to advanced manufacturing, HDF technologies—HDF4, HDF5, HDFView—have helped users manage, share, and make sense of complex data.

This success is no accident. HDF's open, public-domain status has enabled users to read and write HDF data on nearly every computing platform, in almost any programming language, and across various disciplines. It has encouraged experimentation, creativity, and feedback. Bug fixes, feature requests, contributed tools, and novel use cases from thousands of users have shaped HDF into the robust, versatile data foundation it is today.

But today, HDF stands at a crossroads.

HDF Has Grown—But Without a Common Plan

Despite its impact and widespread use, the HDF project has never had a formal governance structure. Its success has been powered by the commitment of a few organizations and individuals, primarily The HDF Group, with essential contributions from the U.S. Department of Energy, NASA, the National Science Foundation, and a few industrial partners. However, this narrow funding and leadership base is not sustainable. The responsibility for such critical infrastructure must not rest on the shoulders of a few.

Why We Must Act—Together

To ensure HDF continues to serve the world's data needs for the next 20 years and beyond, we propose establishing a community-led HDF governance structure. This is not bureaucracy for bureaucracy's sake—it is a necessary step toward sustainability, collaboration, and shared stewardship. Here's why:

- **To broaden the funding base.** HDF is used by millions but funded by few. This imbalance must change. Through a shared governance structure, new contributors—academic, governmental, and industrial—can co-invest in HDF's future.
- **To coordinate and scale development.** The dedicated team at The HDF Group performs exceptional work, but it's essential to recognize that they can't be experts in every field. The increasing demand for their services surpasses what one organization can realistically offer. By embracing a collaborative approach, we can harness diverse talents, experts, and resources, enabling us to tackle everything from bug fixes to innovative features to language bindings.
- **To give users a voice.** Right now, there is no structured way for the community to set priorities or weigh in on the future of HDF. Governance will provide this voice—through a representative Advisory Council and coordinated outreach efforts.
- **To strengthen outreach and support.** As HDF adoption has grown, so has the need for documentation, tutorials, training, and technical support. A better-resourced, better-organized project can meet these needs head-on.

A Call to the Community

We are not starting from scratch. We have a strong foundation. What we need now is to build a shared vision—together. We invite HDF users, developers, stakeholders, and funders to join in shaping a formal governance framework for the project.

If HDF matters to your work—if you rely on it to move science, engineering, products, services, or innovation forward—this is your opportunity to invest in its future. Together, we can create an HDF Project that is resilient, responsive, and ready for what

Parting Thoughts

Language is a tool that builds worlds - or breaks them.

HDF5’s ethos is exploratory and forward-looking, one of humility, openness, and readiness for adaptation.

HDF5 is more than a passive data container but rather an active mediator of scientific communication and knowledge preservation.

Thomas Jefferson Building, interior,
Library of Congress, Washington DC