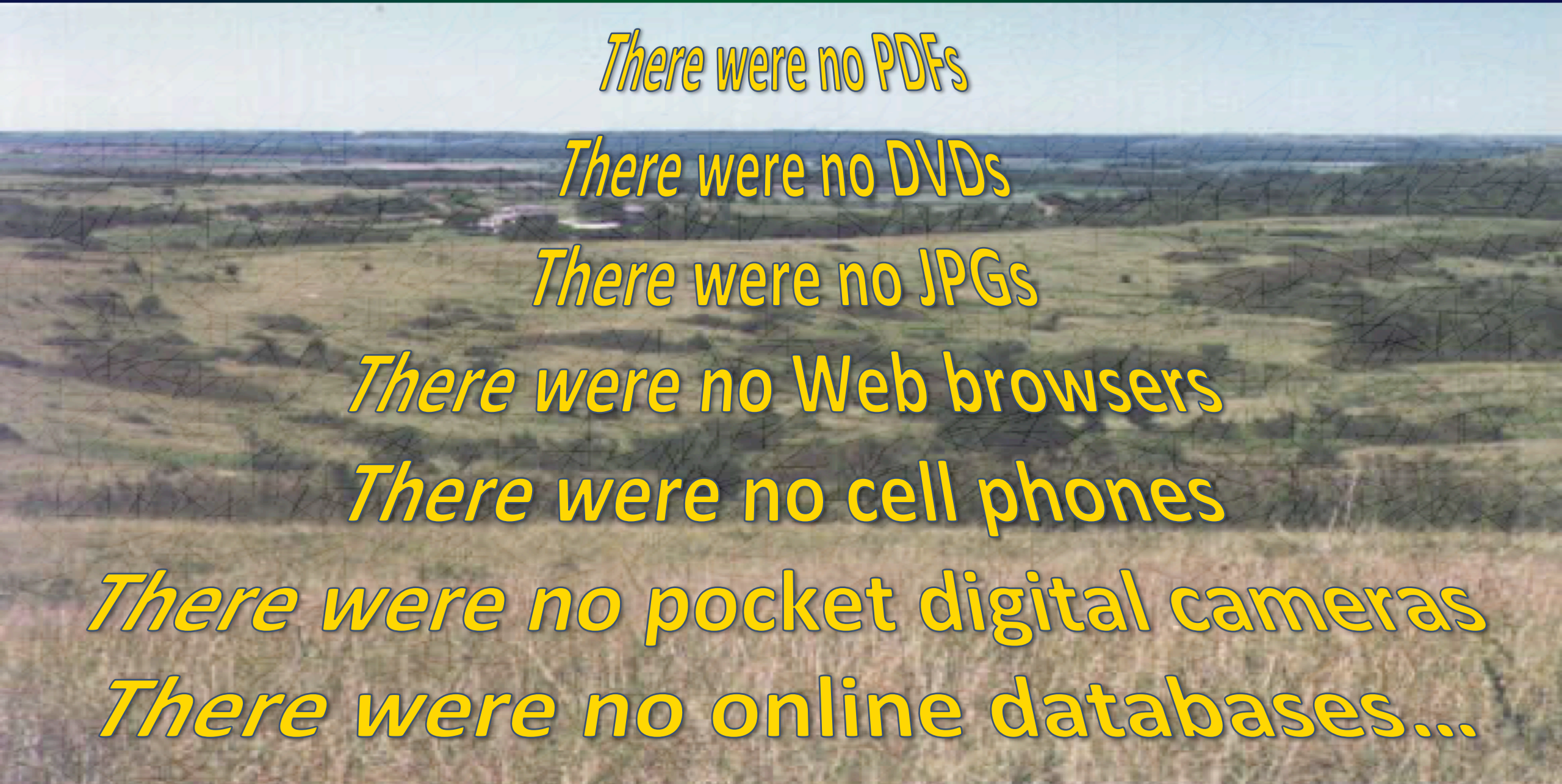
Long, Long ago, on a Prairie Far, Far away….



so we created FIS….

and the world has never been the same.

FIFE Information System Firsts

(with just a little “lying and boasting”)

- First interactive scientific database on the Internet.
- First usable scientific documentation standards for digital data sets.

- First digital scientific data publication to meet the “20-year” test.
- First digital compressed document file format and reader.

The Mission Began Quietly…

In a non-descript warehouse,

just inside the gates of a sleepy government research facility,

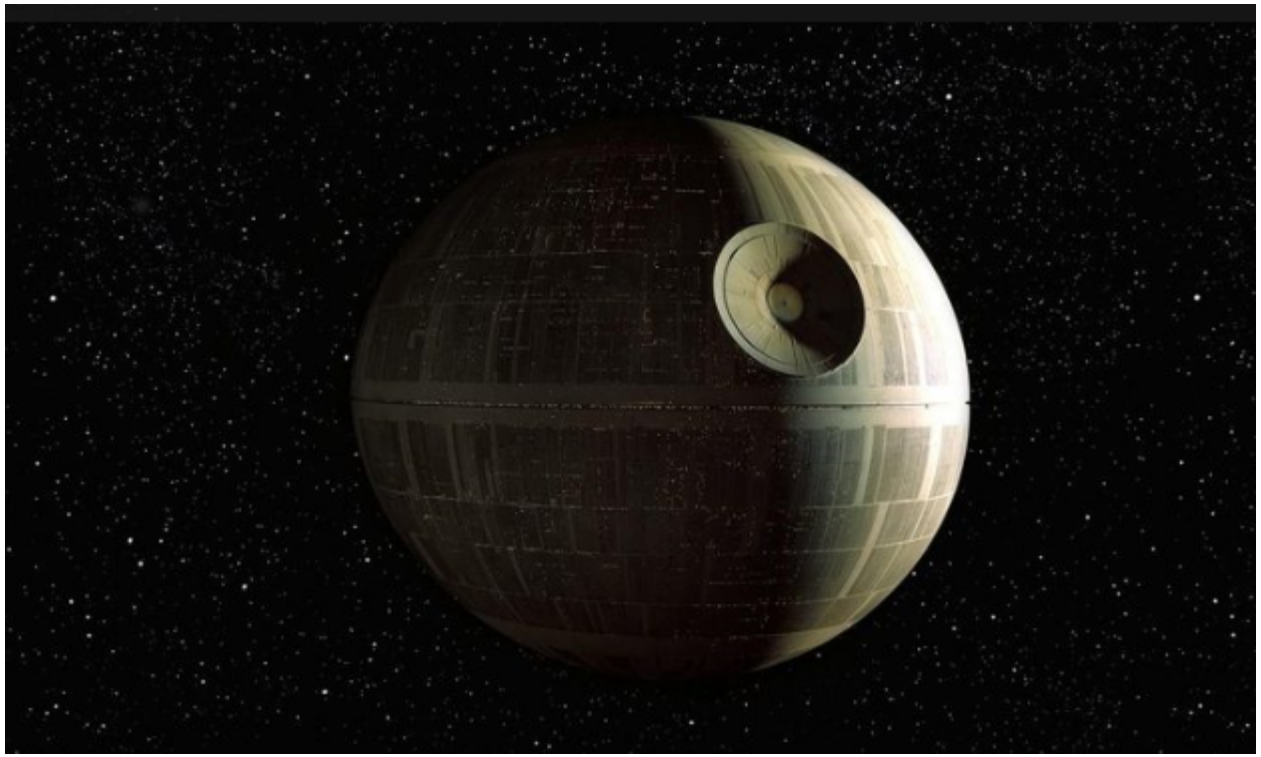
on the outskirts of the capital city.

"Your assignment is to be our eyes and ears and to keep NASA's Data Empire on track as they create a Data System for FIFE," said Forrest, and Piers, and Jim.

"Make sure they understand the requirements and test everything so that it works when we need it. The data from the field will come in every shape and form, in small packets and large, sometimes fast, sometimes slow, but it CANNOT DISAPPEAR INTO A WRITE-ONLY DATA BASE. We must share the data and do Integrative Science!"

But the Empire fought back….

They had plans
They had funds
They had requirements
They had procedures
They had computers
They had pilot systems



Don and Jeff listened, and watched, and tested.
Data went in, and didn't come out.

The Empire had created the ultimate
black hole for data.

To Forrest and Piers and Jim, we reported:

To avoid destruction by the DAAC Star,
FIFE would have to capture the data
before it crossed the event horizon.

OMG….Now what do we do?

“Make it so!” said Forrest and Piers and Jim.

“There are a few old computers and some image processing equipment in the back of the warehouse that you can use.”

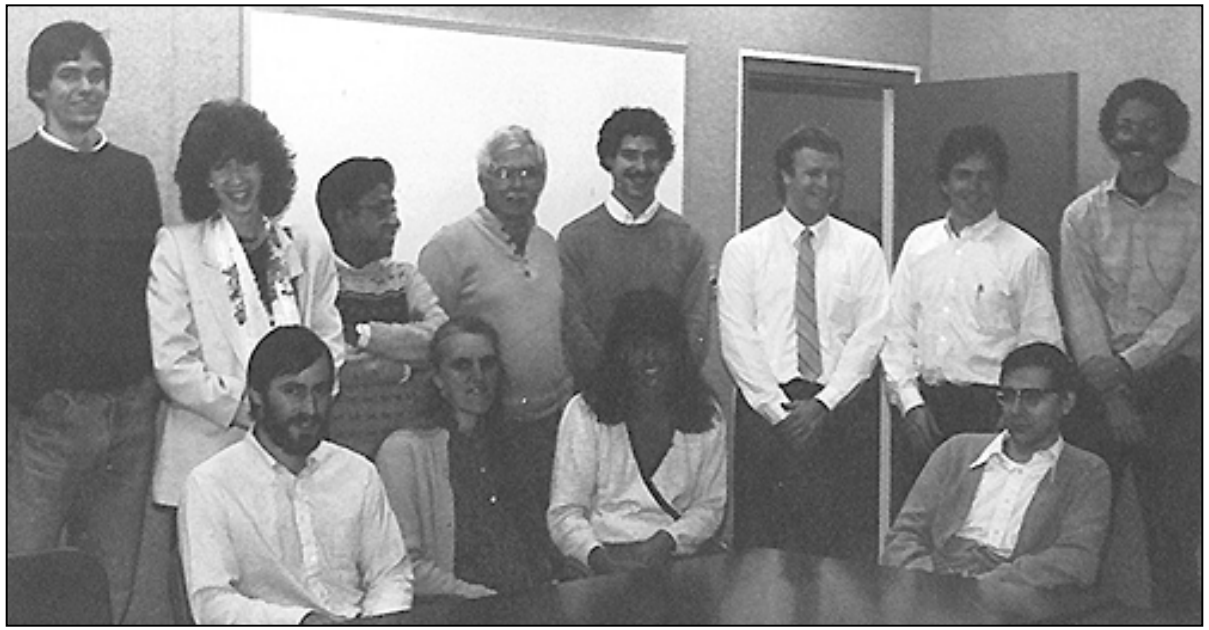
…….and thus was born the FIFE Information System.

We assembled a motley crew….

Over a period of years, before, during, and after the FIFE field seasons, the FIFE Information System Team drew upon the talents of many NASA scientists, contract support scientists, the Konza LTER staff, and PI teams.

Others not shown include Diana Van Elburg-Obler, Fred Huemmrich, Jim Ormsby, Steve Ungar, Jim McManus, Bob Lutz, Alan Nelson, Patrick Agbu, and Blanche Meeson, along with several interns.

(*Apologies to anyone I missed – the computers of that era are long gone and all I have left is human memory and several boxes of paper files. DES)*



FIS Staff, ca 1988. (Standing, left to right) Scott Goetz, Gail Beckman, Babu Banerjee, Forrest Hall, David Landis, Jeff Newcomer, Rich Irish, Tom Eck. (Seated, left to right) Ned Horning, Ruth Kennard, Jaime Nickeson, Don Strebel.

And Started a Revolution….

- We defined what a Scientific Information System should be.
- We focused on capturing and documenting the data in simple standard formats.
- We decided not to spend time and effort developing fancy analysis tools.
- We established (plain text) standards for data files.
- We developed standard image formats and compression standards that worked across computer platforms (JPGs, LZW compression, etc. were still years away).
- We developed documentation standards, before there was "metadata" (based on Diana combining the formats of every lab class she ever took).

We broke a few sacred data system rules…

- The data base was created by working scientists, who were also users.
- We threw out the rigid requirements approach and frequently redesigned the data tables in response to new incoming data or user needs.
- We sent the info system guys into the field to meet the research teams and understand what was happening on the ground.
- Our own use of the system, as scientists, provided additional QA and real-world insights into gaps in data organization or documentation.
- We put users first, e.g. we created a flexible menu system (essentially a primitive web browser) to aid users in finding and downloading FIFE data.
- **And we put it online!** FIS was the first interactive online scientific database that worked, for NASA, for the World.

1986…1987…1988…1989…1990…1991…1992…

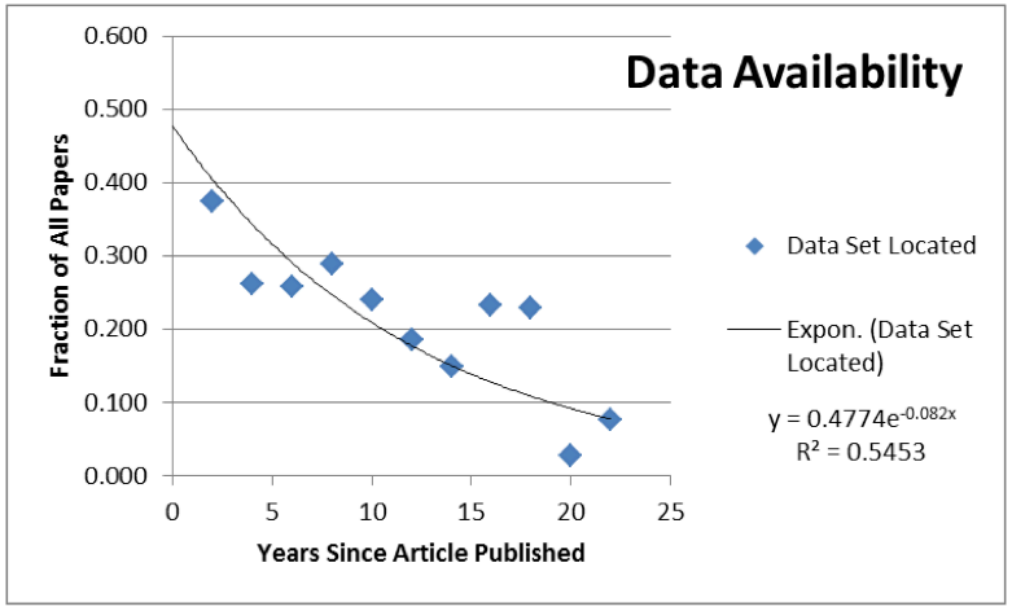
Years passed.

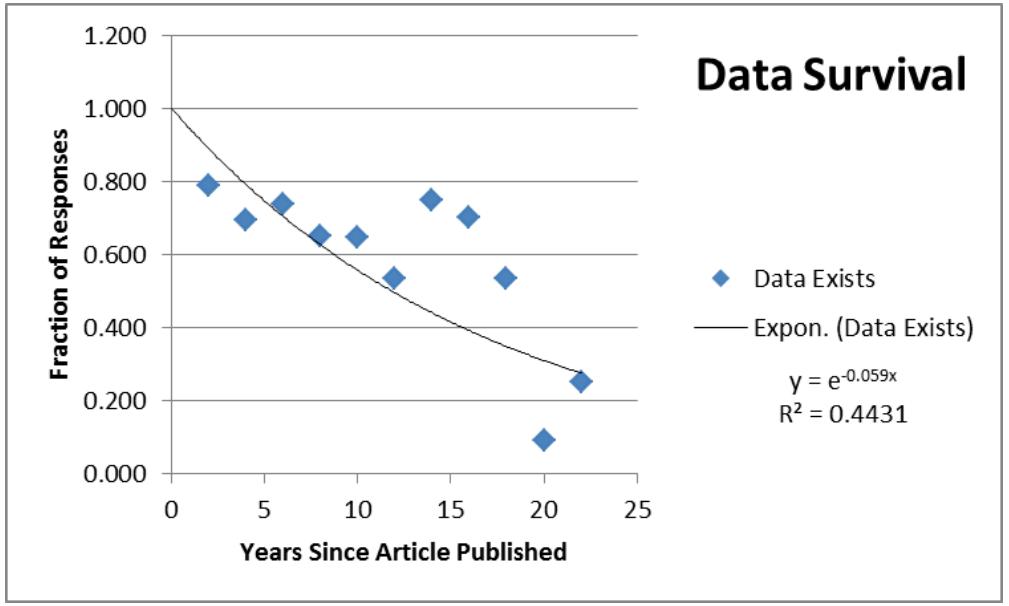
FIS grew into an operational data system that successfully supported FIFE, the FIFE follow-on, and the interdisciplinary investigators during the funded analysis period.

But it wasn't enough…

The FIFE data would be valuable for decades as a baseline for future climatological studies in a changing world. In order for interdisciplinary science to work, many other scientists had to not only have access to the data, but be able to understand how it was collected and what its uses were, for years to come.

The Problem





Access to Unpublished Data Sets Decays Rapidly

A study of 516 published papers without associated published data sets (Vines, et al., 2014) found:

- More than 50% of data sets that were not published and archived were unavailable within a few years. Frequently, contact information was insufficient to locate a responsive data steward.

- Even when data sets had been preserved and stewards were located, the ability to recover the data (viability) deteriorated rapidly. Common reasons were changes in storage technology, currently unsupported storage formats, or documentation not suitable for dissemination.

The longest running FIFE Experiment….

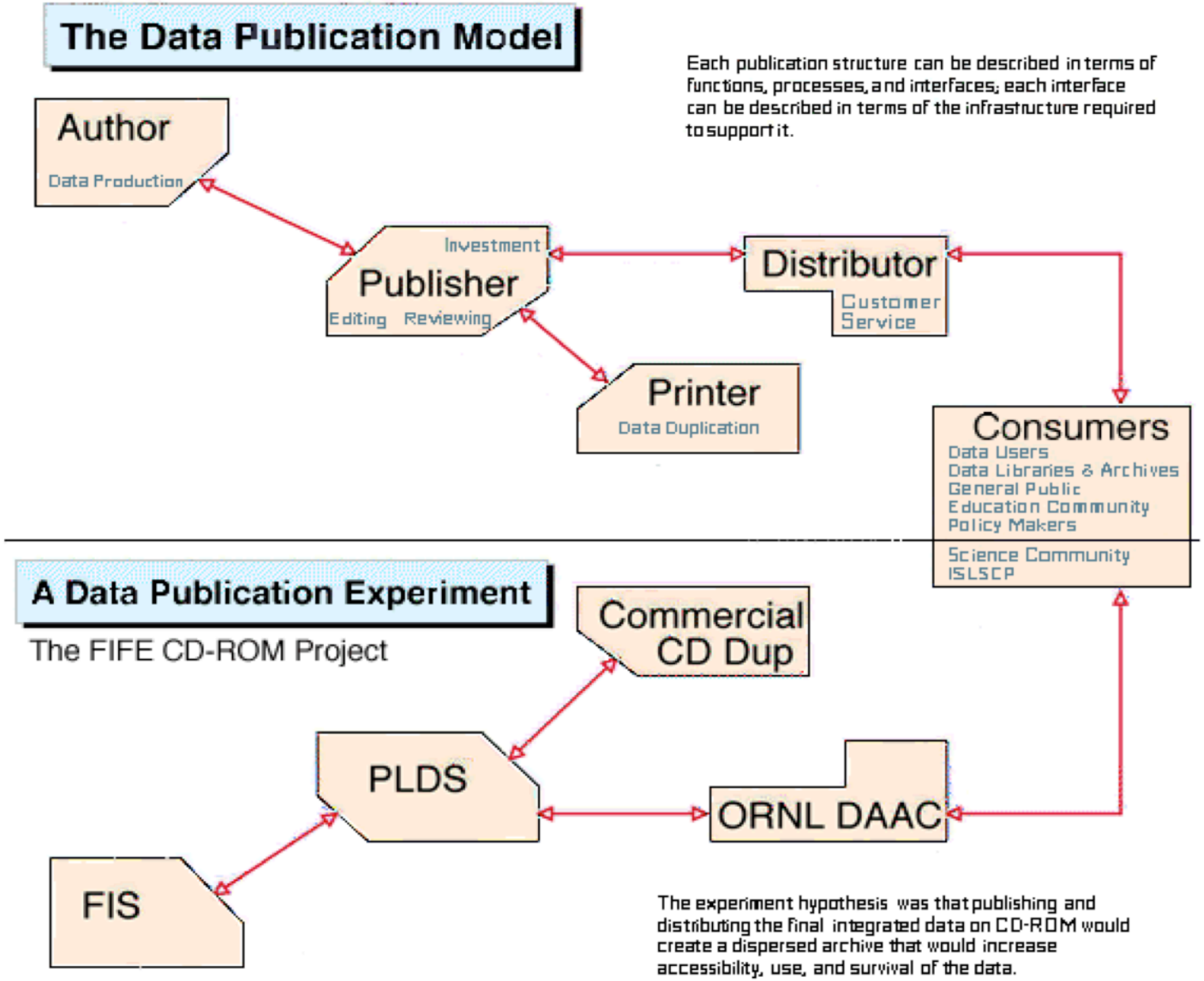
FIS could not be supported beyond the formal end of FIFE.

Hypothesis: Applying a publishing approach to the management of scientific data would ensure that data sets would remain viable for future use by investigators unfamiliar with the original research.

So, we began what is probably the longest running FIFE experiment: **the FIFE data publication study**, which continues to this day.

> *To treat data like a scientific publication, however, we needed to develop all of the tools of publication, including editors and peer reviewers, publishers and disseminators.*

We Created a New Paradigm…



And Executed It Flawlessly….

Published Data Set Collection Details

The body of knowledge captured from FIFE on 5 CD-ROMs amounts to:

- over 100 data sets
- hundreds of compressed image products
- numerous scanned photographs with visual details of the experiment site in Kansas
- extensive documentation of the project and data.

The raw data included more than 1000 distinct parameters and variables and was submitted from approximately 50 independent sources.

Data were stored in simple formats on durable, inexpensive media. Several hundred copies were reproduced and distributed worldwide. Copies were provided to data archives/distribution centers, specifically the Oak Ridge National Laboratory DAAC. The 5 volumes were issued between 1992 and 1994.

The Test…

"Scientific data should be usable after 20 years by a scientist unfamiliar with the data sets or their collection."

Versions of the "20-year test" are widely cited in the data literature and recommended in national guidance documents. For example, the 1991 NRC Report "Solving the Global Change Puzzle" states:

> *Data documentation must pass the well known "20-year test." That is, will someone 20 years from now, not familiar with the data or how they were obtained, be able to find datasets of interest and then fully understand and use the data solely with the aid of the documentation archived with the dataset? [NRC 1991, Cpt. 4, p. 34]*

We are currently (20 years after publication) evaluating how well the FIFE data can be accessed and used by investigators unfamiliar with FIFE or the data.

Four Tests for Key Indicators of Success

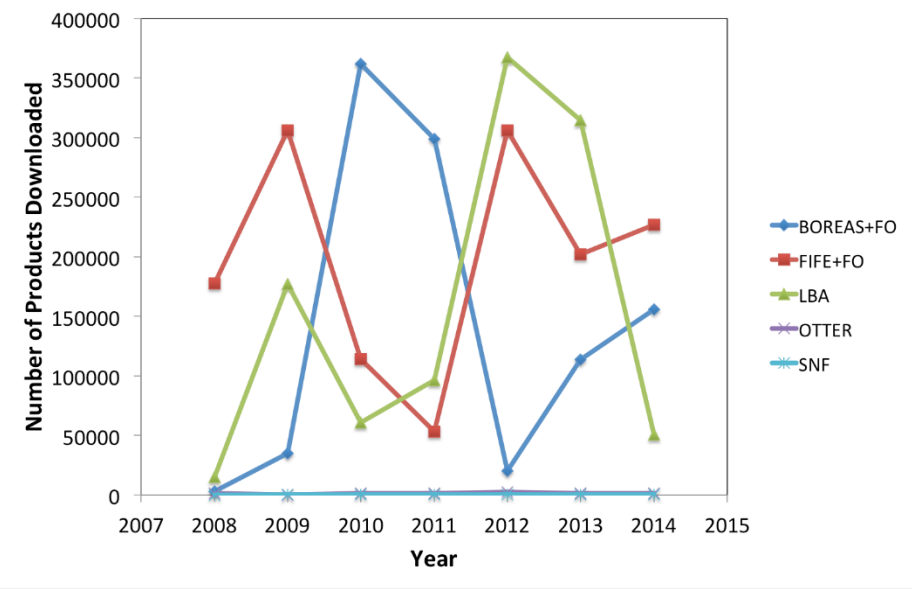
Preservation: The data exist on CD-ROMs and in an online archive.

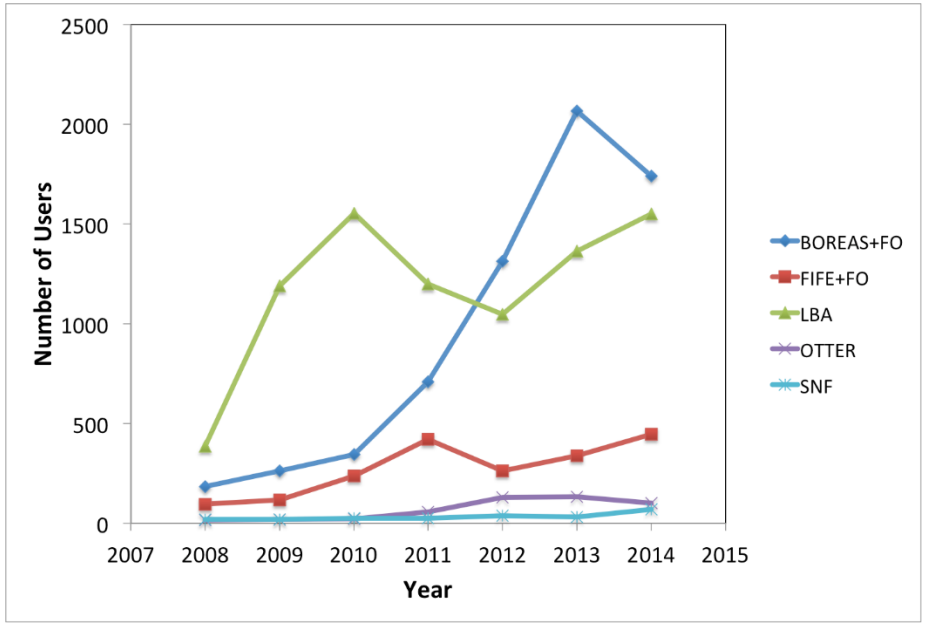
Accessibility: Worldwide Internet access to the online archive; CD-ROMs are readable on all current computers.

Viability: Data files import directly into spreadsheets; still evaluating ability to recover image files and other compressed formats from CD-ROMs.

Vitality: Data are still being used in current scientific papers.

The Result…



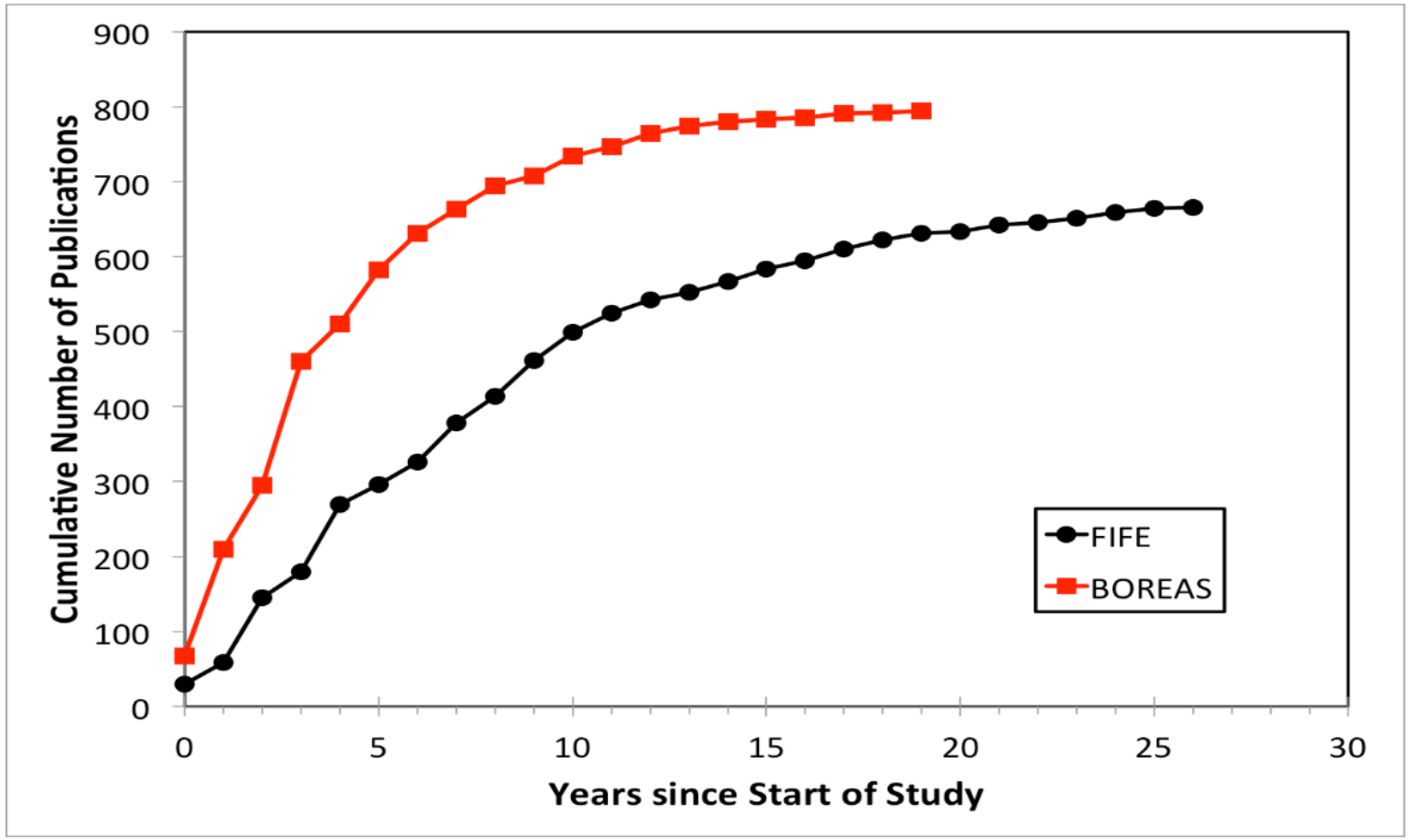


- The published data sets from FIFE, and from BOREAS, were distributed both by CD-ROM and online by the ORNL DAAC.
- Online access statistics indicate number of unique users ramped up quickly and has continued at 100 to 300 users per year for FIFE.
- And has grown to over 2000 for BOREAS.

(ORNL Data provided by R. Cook and B. McMurray, ORNL DAAC)

No Data Decay Curve!

FIFE and BOREAS data continue to be used in scientific papers 20 years after the publication of the data sets.



Publication is a Proven Solution for Information Preservation

- "Publication" in the sense of stand-alone copies of scientific works has been successful over centuries and cultures. Parchment manuscripts from the 9th Century still exist today, indicating a survival probability of 0.85 per century (Cisne, 2005). A modest initial number of copies, or regular reproduction of more copies, is sufficient to ensure long-term survival.

- The publication model for data is to produce documented, stand-alone archives on distributable media, implementing procedures and processes analogous to those that work for books.

- The medium or distribution mechanisms may change, e.g the FIFE CDs are now on the Internet, but the fundamental result of a publication effort is strong enough to survive such technological change.

What We Learned…

Lessons Learned

- It is important to have a dedicated data staff with scientific training to prepare data for publication.
- Data field names, units, formats, and descriptions must be standardized, and consistent among data sets.
- Investigators must be in the loop to provide feedback during QA, data integration, cross-checking, and pre-publication review.
- People issues are as important as technical issues, especially with respect to data submission, quality assurance, and peer review.

The Scientific Paper Analogy for detailed data set documentation:

1. Title
2. Investigators
3. Abstract/Introduction
4. Theory of Measurements
5. Equipment
6. Procedure
7. Observations
8. Data Description/Glossary
9. Data Manipulations
10. Errors
11. Notes
12. References
13. Data Access
14. Output Products & Availability
15. Glossary of Terms

Best Practices

- Preserving data for future use, even your own data, should be treated as a data publication project to ensure preservation.
- Focusing on data integration while preparing for publication is crucial to ensure data vitality.
- Data sets must be clearly and completely documented, as if writing a scientific paper about the data.
- Data and documentation must be kept together to ensure data set viability.
- Use standard formats and long-term media; try to anticipate changing technology, loss of funding for archives or data migration.
- Choosing a suitable archive is important to both preservation and accessibility. Do not depend on the Cloud for archiving.
- Educate future scientists in the principles and practices of data publication.

The Final Word

- In FIS, we laid out key concepts that led to the BOREAS Info System (BORIS) and were ported into a number of other developing data systems at the time, including subsequent field experiments throughout the world, the EPA's EMAP Information System, and NFS's LTER Information Network.

- Some of these innovations and concepts are still visible in the ORNL DAAC, which eventually inherited the FIFE and BOREAS data.

- As we look around at scientific data systems today, while the technology and standards have improved, developers still struggle with some of the same issues, which we believe are "people" issues and will always be hard to automate:
  - Appropriate quality checking of data.
  - Obtaining complete and readable documentation.
  - Data integration across diverse data sets, time, and space
  - Anticipating and overcoming technological change and evolution of science standards and approaches.