Perspectives on Research Problems in Family History from the LDS Family and Church History Department

April 3, 2003

Future Directions in Family History

- Concentrated effort to make family history easier for the non-genealogist
  - Common Pedigree
    - Single interface
    - Detect matches
    - Enable collaboration
  - FH Research
    - Simpler research model for non-genealogists
    - World record manager
    - Online images

Family History Research is Exciting!

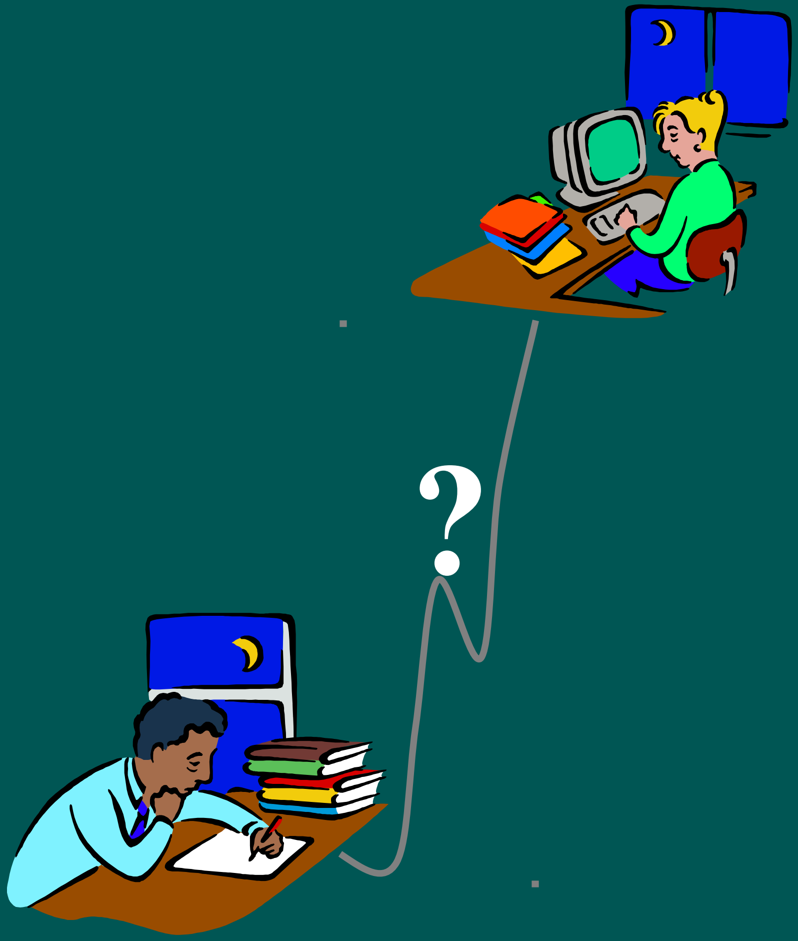
- Research problems exist in many areas of computer-science and engineering
- Problems are quite challenging and have broad application
- Millions of people who struggle to provide saving ordinances for their ancestors would benefit

Research Problems

- Common Pedigree
  - Record linkage
  - Data standardization
  - Efficient data access
  - Expert finding
- FH Research
  - Faster image indexing
  - Digital image delivery
  - Digital image conversion and storage
  - Image enhancement
  - Context-sensitive help
  - Catalog-data extraction
  - Language translation
  - Indexing external data
  - Digital data preservation
  - Future digital data access

Record Linkage

- Given two people in two different pedigrees, are they really the same person?
  - Common problem in census analysis, healthcare
  - Rules vs. statistical models
    - Training data vs. statistical model vs. combination
- Given a person in a pedigree and a large set of genealogical records, do any of the records match?



Data Standardization

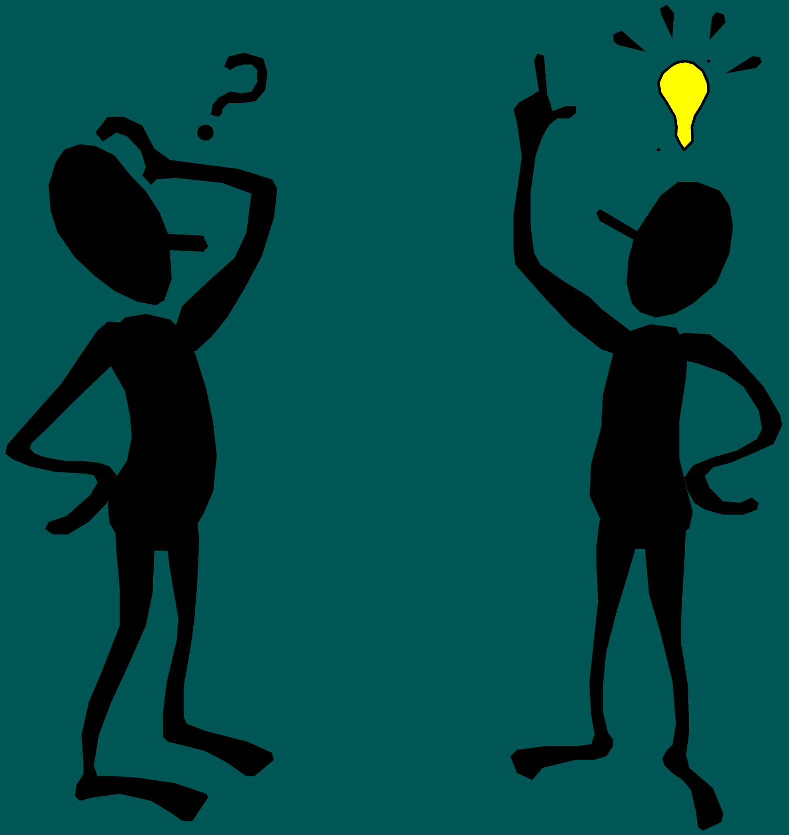
- Good standardization essential for record linkage
  - Henry Thomas = Hank Thomas = Hank Tomas
  - Thomas Henry = T?om Henry = Tom H?anks
- Similar person-names?
  - Requires name-parsing (Rules vs. HMMs)
- Nearby locales
  - Analyze migration patterns?
- Another idea: shared acquaintances
  - Look at close neighbors or document witnesses?

Efficient Data Access

- A single pedigree/descendency screen could display 30-60 people
- Each person may require reading 10 database records
- For every new person entered, we need to find potential matches – Requires complex queries
- Possible solutions:
  - Distributed cache?
    - Need to cluster and balance objects in each partition
    - Twist on traditional object caching: intensional cache description
  - Peer-to-peer?

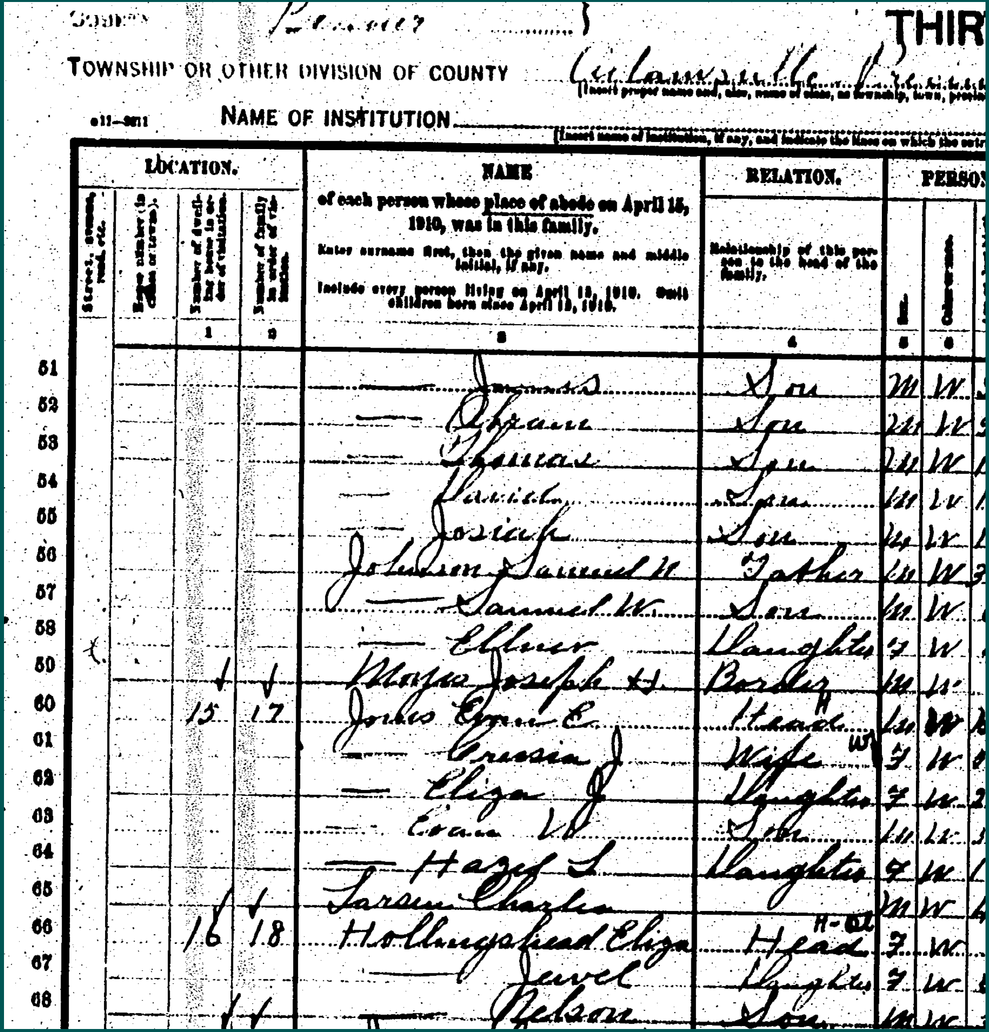
Expert Finding

- General problem is well-known
  - Tacit Knowledge Systems, Autonomy
  - Analyze email and documents to identify key terms related to an individual
- Unique aspects of FH
  - Watch tasks, not keywords
  - Determine whether someone is “good” at performing those tasks



Faster Image Indexing

- People currently index images manually
- Desired approach:
  - Two independent indexers + adjudication
- Four problems:
  - Identify field boundaries
  - Recognize handwriting
  - Verify human indexing results
  - Find matches without indexing



COUNTY [illegible]

THIR

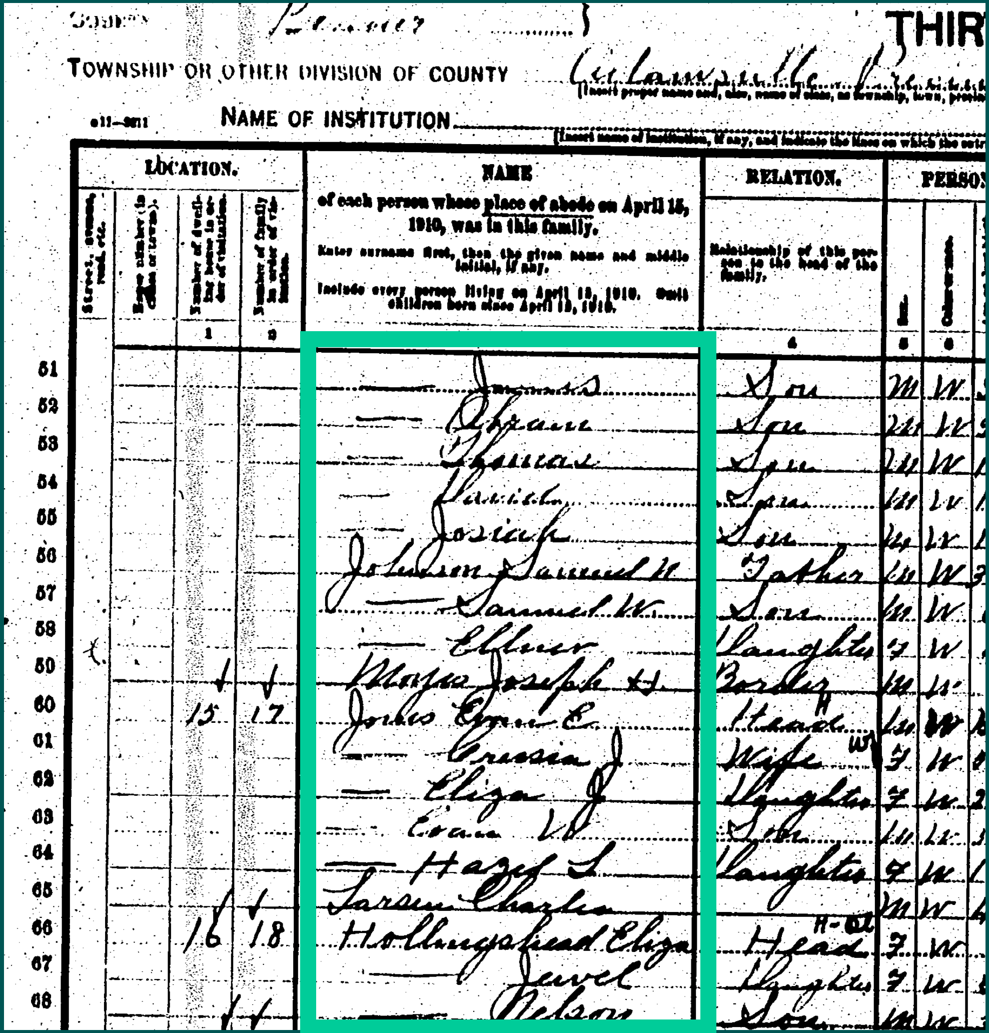
TOWNSHIP OR OTHER DIVISION OF COUNTY [illegible]
[Insert proper name and, also, name of class, as township, town, precin

NAME OF INSTITUTION
[Insert name of institution, if any, and indicate the lines on which the entr

	LOCATION.				NAME	RELATION.	PERSO	
	Street, avenue, road, etc.	House number (in cities or towns).	Number of dwelling house in order of visitation.	Number of family in order of visitation.	of each person whose place of abode on April 15, 1910, was in this family. Enter surname first, then the given name and middle initial, if any. Include every person living on April 15, 1910. Omit children born since April 15, 1910.	Relationship of this person to the head of the family.	Sex.	Color or race.
			1	2	3	4	5	6
51					— James S	Son	M	W
52					— Ephraim	Son	M	W
53					— Thomas	Son	M	W
54					— Daniel	Son	M	W
55					— Josiah	Son	M	W
56					Johnson Samuel H	Father	M	W
57					— Samuel W.	Son	M	W
58					— Elmer	Daughter	F	W
59					Moyer Joseph H	Border	M	W
60			15	17	Jones Lynn E	Head	M	W
61					— Orissia J	Wife	F	W
62					— Eliza J	Daughter	F	W
63					— Evan W	Son	M	W
64					— Hazel L	Daughter	F	W
65					Larson Charles		M	W
66			16	18	Hollingshead Eliza	Head	F	W
67					— Jewel	Daughter	F	W
68					— Nelson	Son	M	W

Digital Image Delivery

- Can we deliver readable images over a 28K line?
  - Targeting
  - Compression
- Needed for indexing as well as original image lookup



COUNTY Preston

THIR

TOWNSHIP OR OTHER DIVISION OF COUNTY [illegible]

[Insert proper name and, also, name of class, as township, town, precinct

NAME OF INSTITUTION

[Insert name of institution, if any, and indicate the lines on which the entr

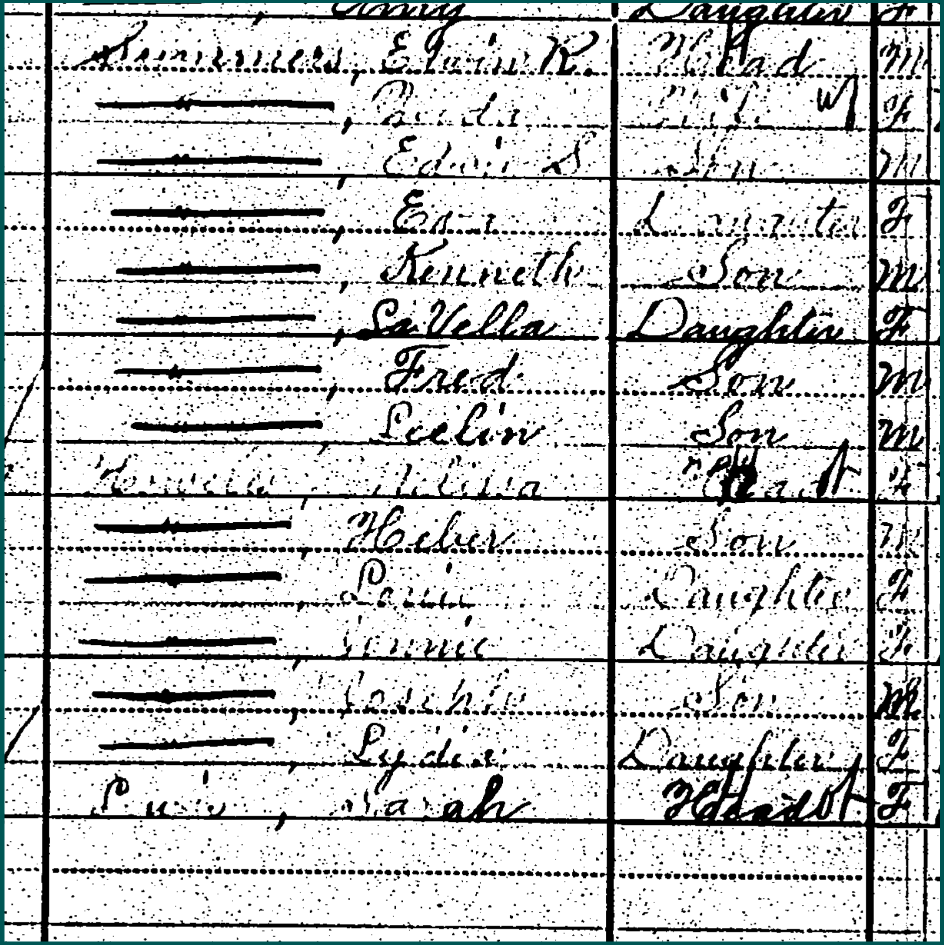
	LOCATION.				NAME	RELATION.	PERSON	
	Street, avenue, road, etc.	House number (in cities or towns).	Number of dwelling house in order of visitation.	Number of family in order of visitation.	of each person whose place of abode on April 15, 1910, was in this family. Enter surname first, then the given name and middle initial, if any. Include every person living on April 15, 1910. Omit children born since April 15, 1910.	Relationship of this person to the head of the family.	Sex.	Color or race.
			1	2	3	4	5	6
51					— James	Son	M	W
52					— Abram	Son	M	W
53					— Thomas	Son	M	W
54					— Daniel	Son	M	W
55					— Josiah	Son	M	W
56					Johnson Samuel H	Father	M	W
57					— Samuel W.	Son	M	W
58					— Ellner	Daughter	F	W
59			/	/	Moyer Joseph H	Boarder	M	W
60			15	17	Jones Ezra E.	Head	M	W
61					— Cressia J	Wife W	F	W
62					— Eliza J	Daughter	F	W
63					Evan W	Son	M	W
64					— Hazel L	Daughter	F	W
65			/	/	Sarum Charles		M	W
66			16	18	Hollingshead Eliza	Head H-O	F	W
67					— Jewel	Daughter	F	W
68			/	/	— Nelson	Son	M	W

Digital Image Conversion and Storage

- If we were to convert all of our 2.2M rolls of microfilm to digital images:
  - At one roll per hour, 24 hours per day, 6 days per week, it would take 300 years
  - At 2 Mb per image, it would occupy >2 Pb
- Of course, wouldn't convert everything right away, if ever
  - 50% of requests are for <5% of films
  - 5% of films would require 100 Tb and 15 years
- Possible solutions
  - "Ribbon" scanning?
  - Hierarchical and/or distributed storage?

Image Enhancement

- Image enhancement is a well-known problem
- Does knowing the type of information to expect make it any easier?



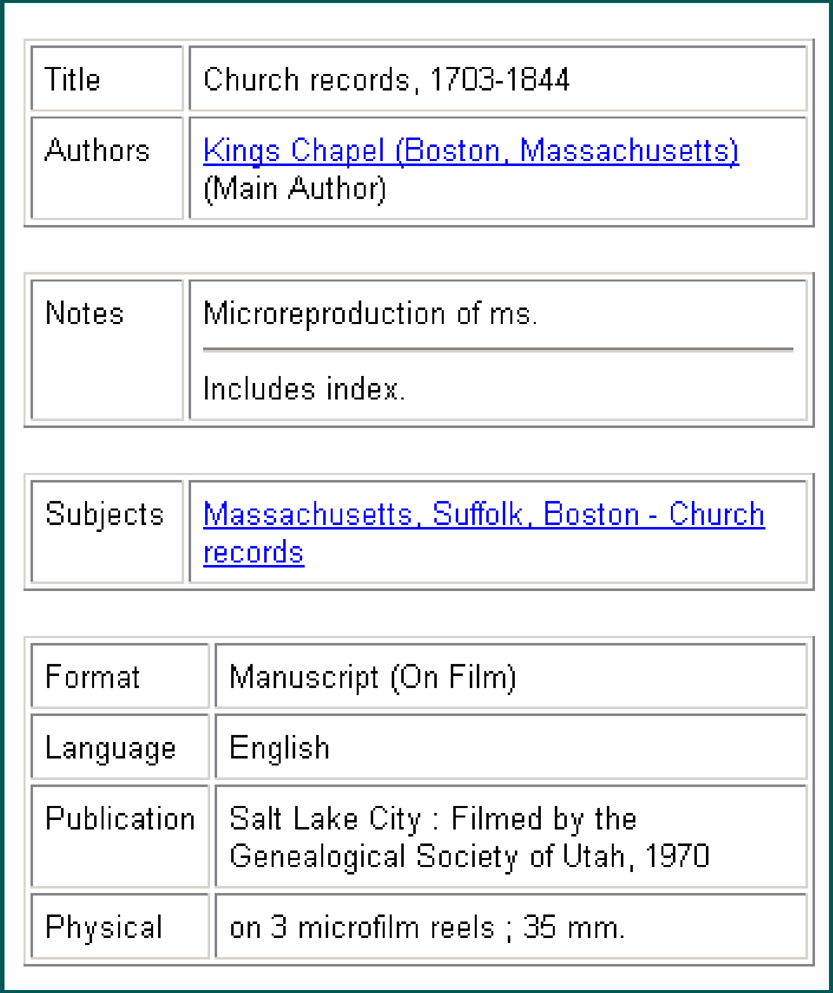
[illegible]	Daughter	F
Summers, Edwin R.	Head	M
—, [illegible]	Wife	F
—, Edwin S.	Son	M
—, Eva	Daughter	F
—, Kenneth	Son	M
—, LaVella	Daughter	F
—, Fred	Son	M
—, Leelin	Son	M
[illegible], [illegible]	Head	F
—, Heber	Son	M
—, Lorin	Daughter	F
—, Jennie	Daughter	F
—, [illegible]	Son	M
—, Lydia	Daughter	F
[illegible], Sarah	Head	F

Context-Sensitive Help

- Goal: help people know what they should do next, and guide them in doing it
  - Help-desk functionality: Question-Answer, Problem-Resolution
  - Task-oriented functionality (TurboTax)
- Can we build the help system collaboratively from patron emails, submissions, etc.?
  - Growing database of questions and answers
  - Flowcharts that transform over time

Catalog-Data Extraction

Catalog Entry



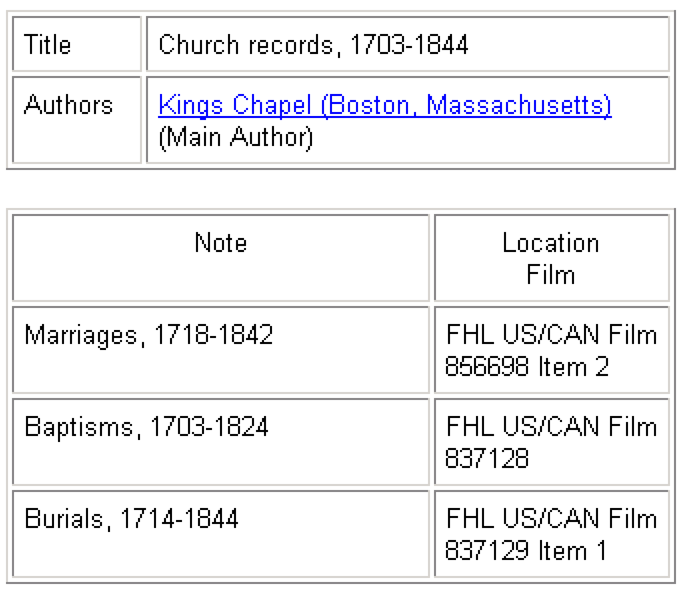
Title	Church records, 1703-1844
Authors	Kings Chapel (Boston, Massachusetts) (Main Author)

Notes	Microreproduction of ms. Includes index.

Subjects	Massachusetts, Suffolk, Boston - Church records

Format	Manuscript (On Film)
Language	English
Publication	Salt Lake City : Filmed by the Genealogical Society of Utah, 1970
Physical	on 3 microfilm reels ; 35 mm.

Film Notes



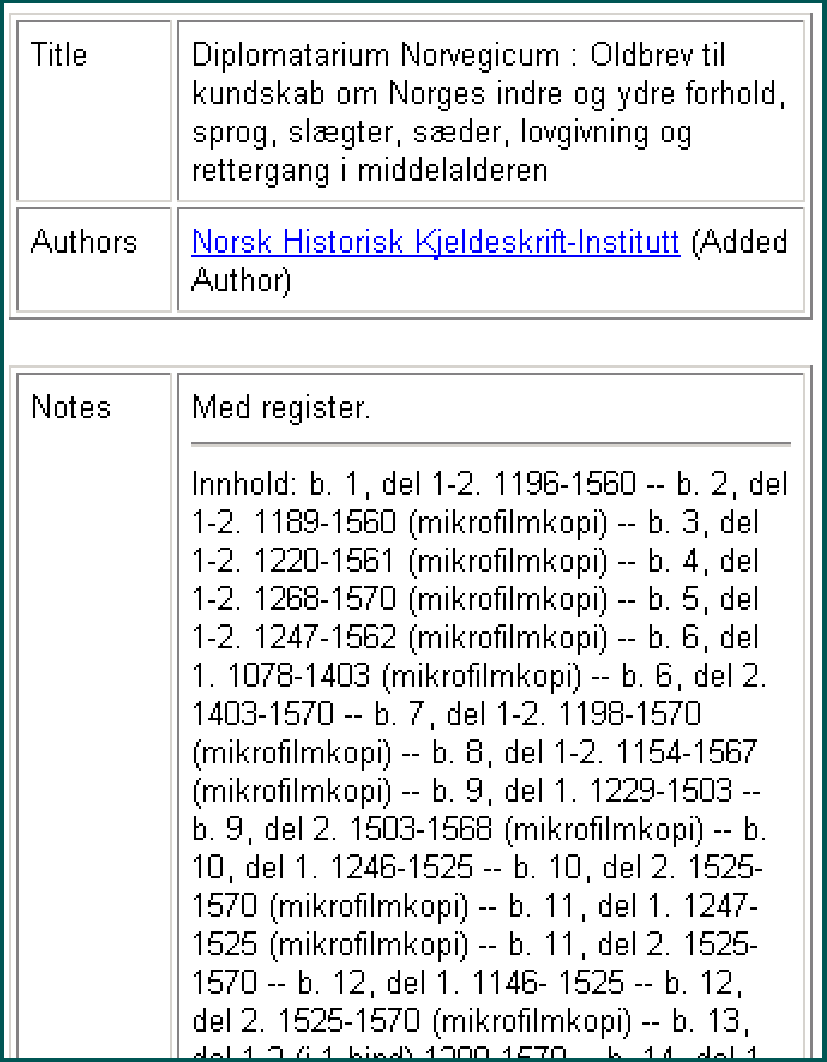
Title	Church records, 1703-1844
Authors	Kings Chapel (Boston, Massachusetts) (Main Author)

Note	Location Film
Marriages, 1718-1842	FHL US/CAN Film 856698 Item 2
Baptisms, 1703-1824	FHL US/CAN Film 837128
Burials, 1714-1844	FHL US/CAN Film 837129 Item 1

Need to extract text into individual fields for improved search!

Language Translation

- Surprisingly, some people can no longer understand the language of their ancestors
- Language translation is simplified due to a known domain and a restricted vocabulary



Title	Diplomatarium Norvegicum : Oldbrev til kundskab om Norges indre og ydre forhold, sprog, slægter, sæder, lovgivning og rettergang i middelalderen
Authors	Norsk Historisk Kjeldeskrift-Institutt (Added Author)

Notes	Med register. Innhold: b. 1, del 1-2. 1196-1560 -- b. 2, del 1-2. 1189-1560 (mikrofilmkopi) -- b. 3, del 1-2. 1220-1561 (mikrofilmkopi) -- b. 4, del 1-2. 1268-1570 (mikrofilmkopi) -- b. 5, del 1-2. 1247-1562 (mikrofilmkopi) -- b. 6, del 1. 1078-1403 (mikrofilmkopi) -- b. 6, del 2. 1403-1570 -- b. 7, del 1-2. 1198-1570 (mikrofilmkopi) -- b. 8, del 1-2. 1154-1567 (mikrofilmkopi) -- b. 9, del 1. 1229-1503 -- b. 9, del 2. 1503-1568 (mikrofilmkopi) -- b. 10, del 1. 1246-1525 -- b. 10, del 2. 1525-1570 (mikrofilmkopi) -- b. 11, del 1. 1247-1525 (mikrofilmkopi) -- b. 11, del 2. 1525-1570 -- b. 12, del 1. 1146- 1525 -- b. 12, del 2. 1525-1570 (mikrofilmkopi) -- b. 13, del 1-2 (i 1 bind) 1200-1570 -- b. 14, del 1

Indexing External Data

- Much more information relevant to FH research information lies outside the LDS Church's holdings than within it
  - Most people stop if the Church can't point them to the information they need
- On the Web
  - Classifying websites, filling out forms, identifying names, dates, places, and record types
- In external databases
  - Mapping and restructuring information from one schema to another

Digital Data Preservation

- Big concern
  - Microfilm lasts 100’s of years, CD’s, DVD’s, and hard disks much less
- Approaches
  - Technical preservation
  - Emulation
  - Migration
  - Convert to analog
  - LOCKSS (Lots of Copies Keeps Stuff Safe)

Future Digital Data Access

- Related to digital data preservation
- Many records offices have switched to storing digital data only – getting rid of paper
- We are usually restricted from accessing their records for 70-110 years
- How can we ensure that we’ll be able to read the digital data that’s being created today, 100 years from now?

Conclusion

- Wide variety of research problems
  - Extremely interesting!
  - Beneficial to mankind!
- We are currently investigating ways to work with people at BYU and others who would like to help with research in these areas

  Contact: Dallan Quass (quassdw at ldschurch.org)
- We are recruiting qualified software engineers

  Contact: Daniel Bray (brayde at ldschurch.org)