Computational Analysis and Visualization of Electronic Records Collections

Visualization and Data Analysis Group
Texas Advanced Computing Center, UT Austin
Research Forum SAA, 2009

Research Motivation

Digital tools for archivists and archives' users to:

Make sense, explore, appraise, describe, discover, manage, and preserve electronic records collections of varied structures and formats

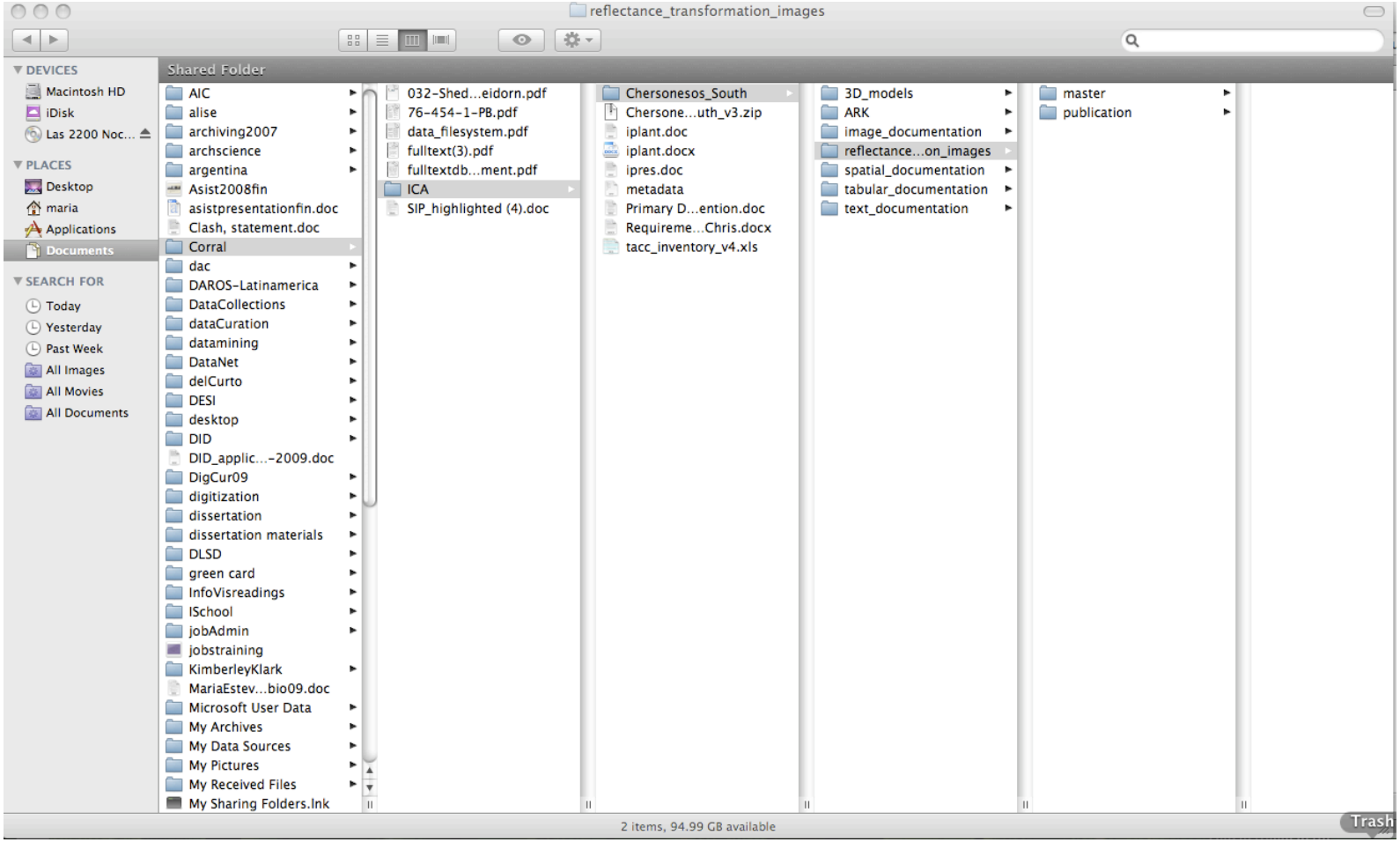
Challenges

- Large archival data
- Minimal metadata
- What metadata can be extracted from electronic records collections?
- What can be inferred from the structure and the content of electronic records records?
- Abstract representations of collections: visual literacy development

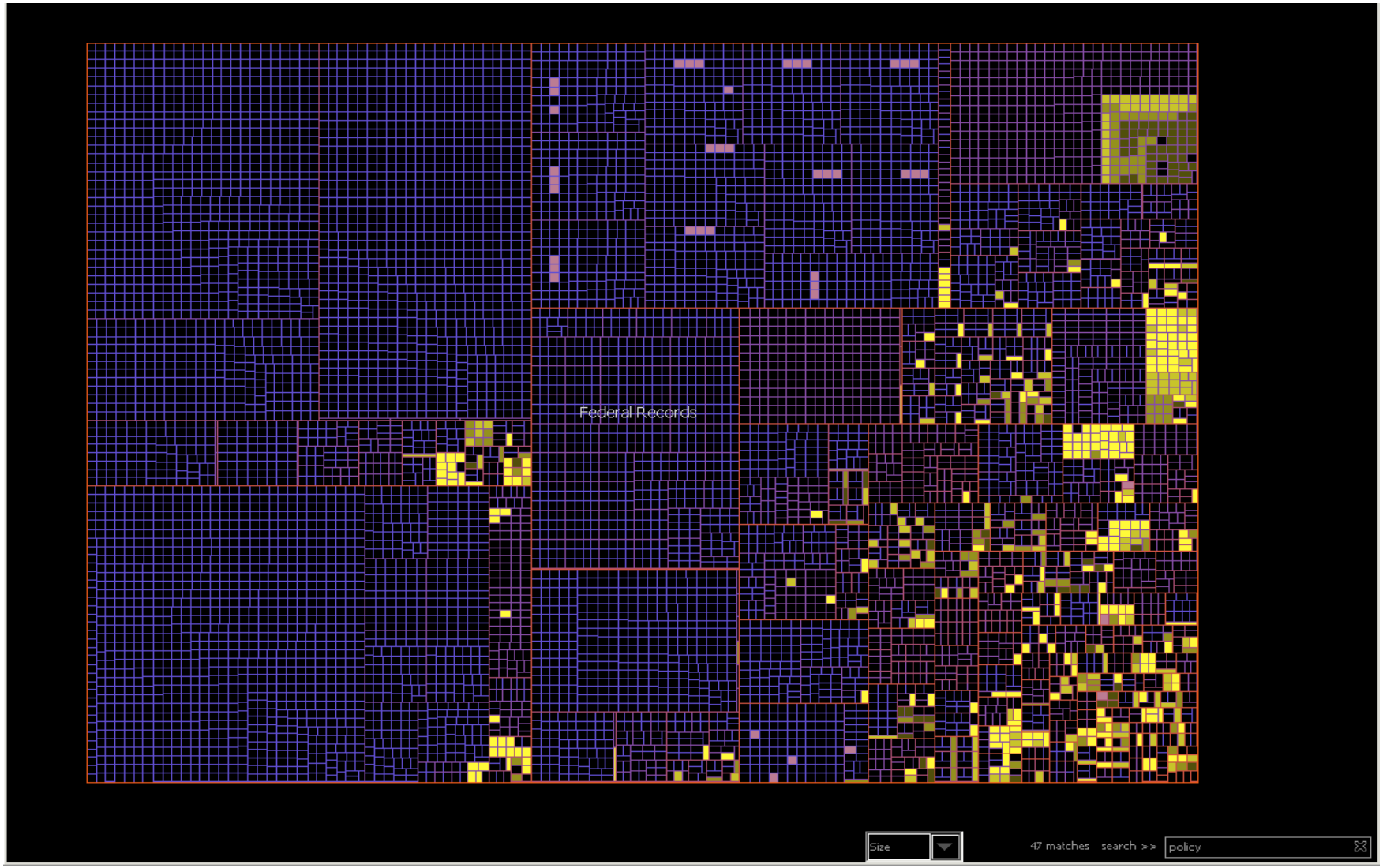
I. Treemaps

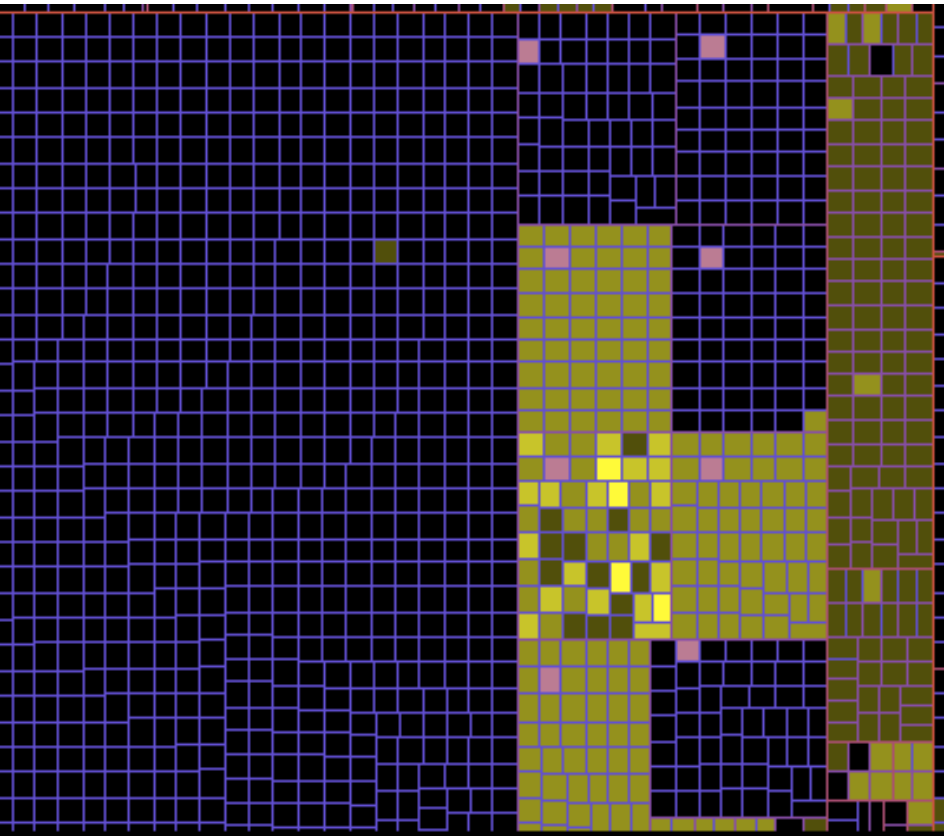
- Developed by Ben Shneiderman in the 1990's
- Adapted to visually scan large collections and to focus on smaller parts
- Collection properties are extracted and stored in a database
  - Structural, descriptive, technical
- Rendered through a treemap visualization application

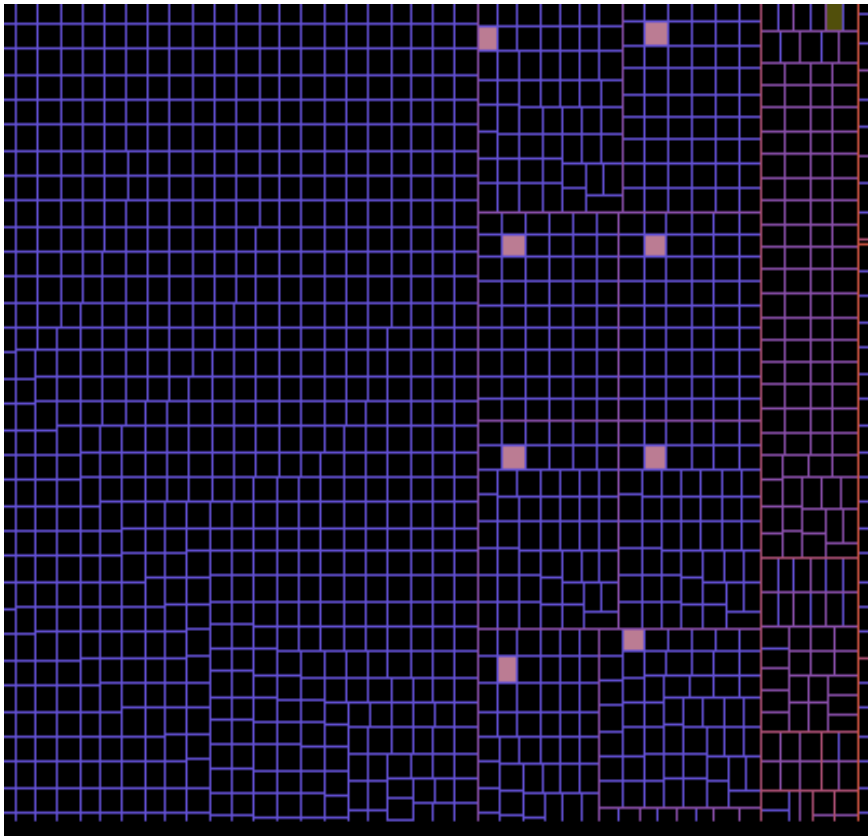
Directories view



NARA test-bed collections in the Transcontinental Persistent Archives Prototype







Records Group: Records of the Bureau of Census

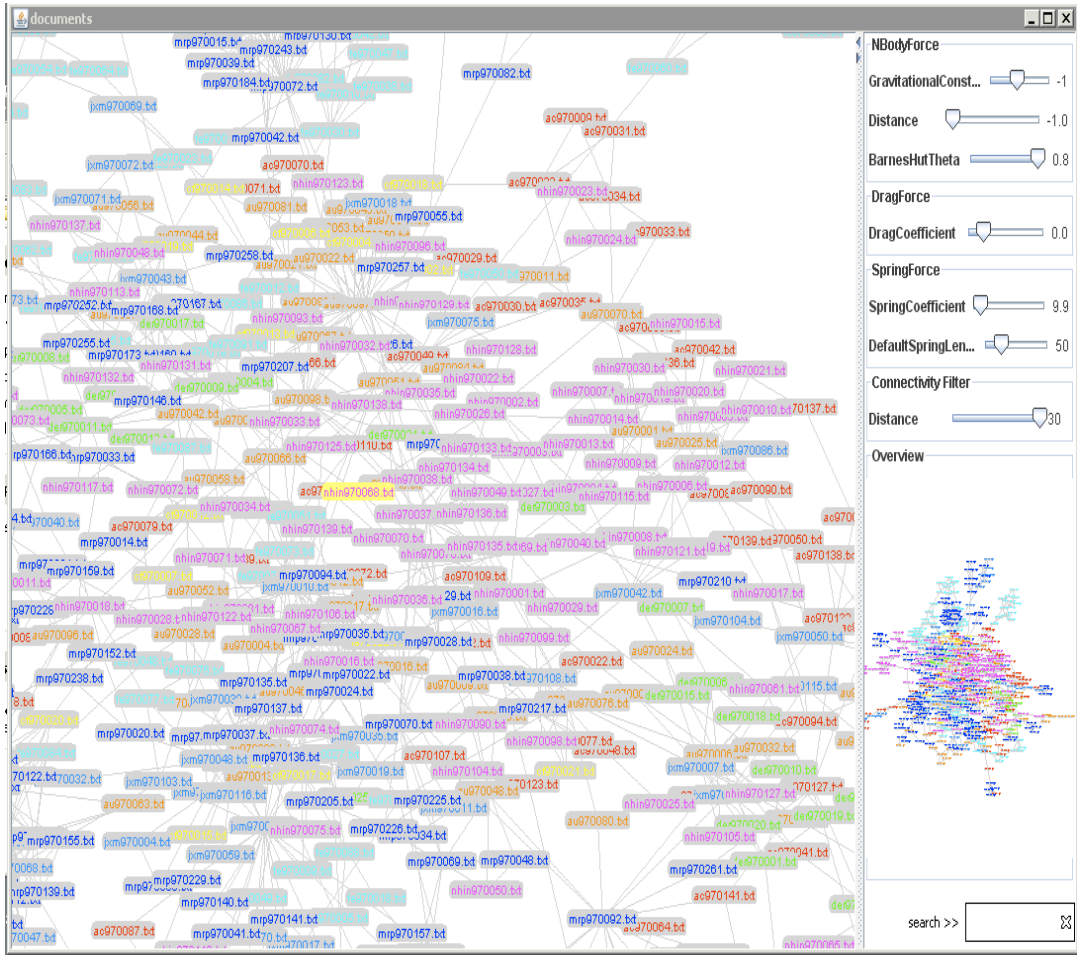
Smallest partition: one directory

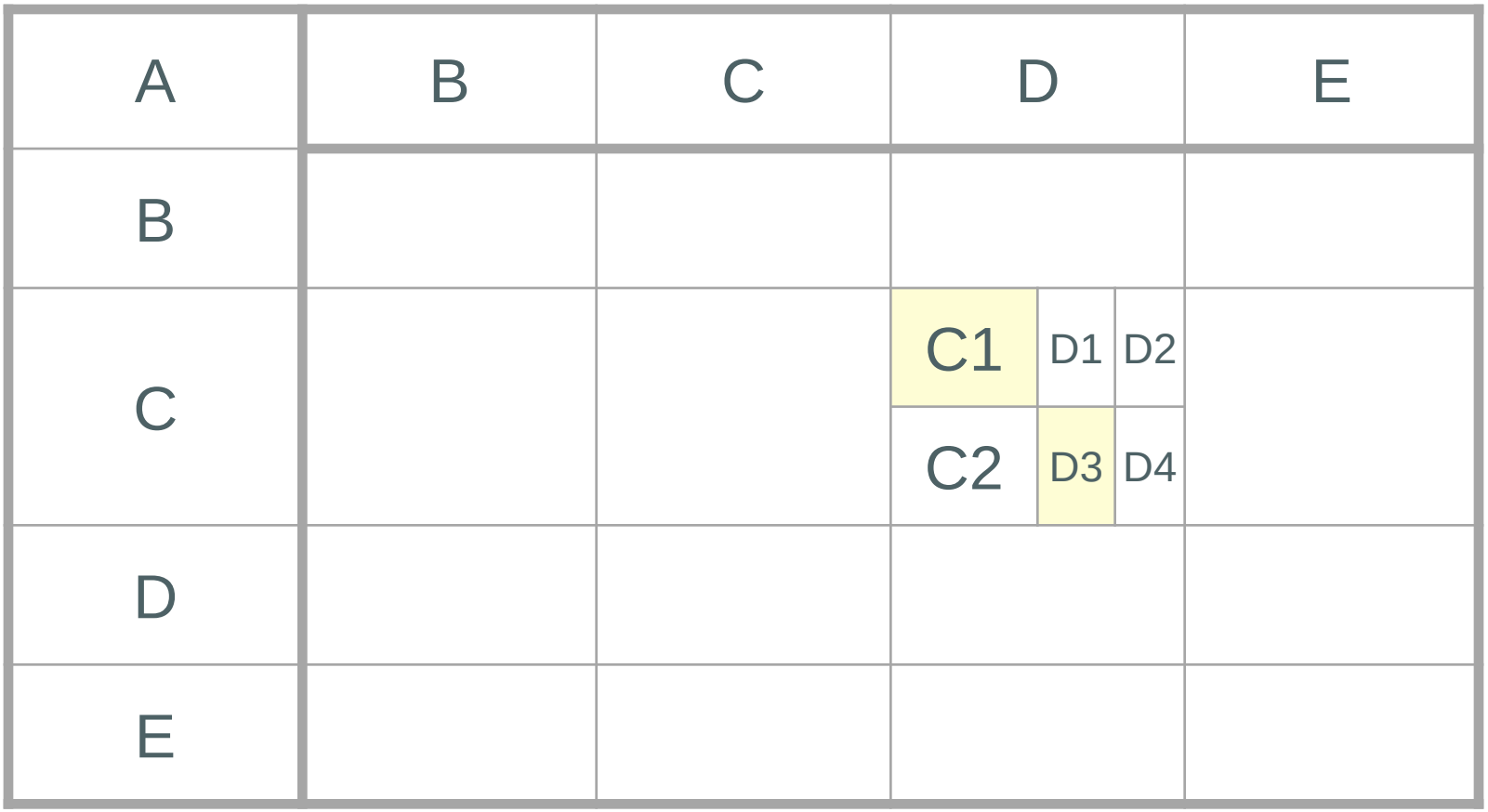
Pink fill: searched keyword (2004) present in the name of the directory

Degrees of yellow-green-brown: from more to less number of files present in each directory / from more to less number of different file extensions present in each directory

II. Paragraph alignment visualization

- Identify related records
  - Belonging to same activities, projects, transactions, events, etc.
  - Content based relationships

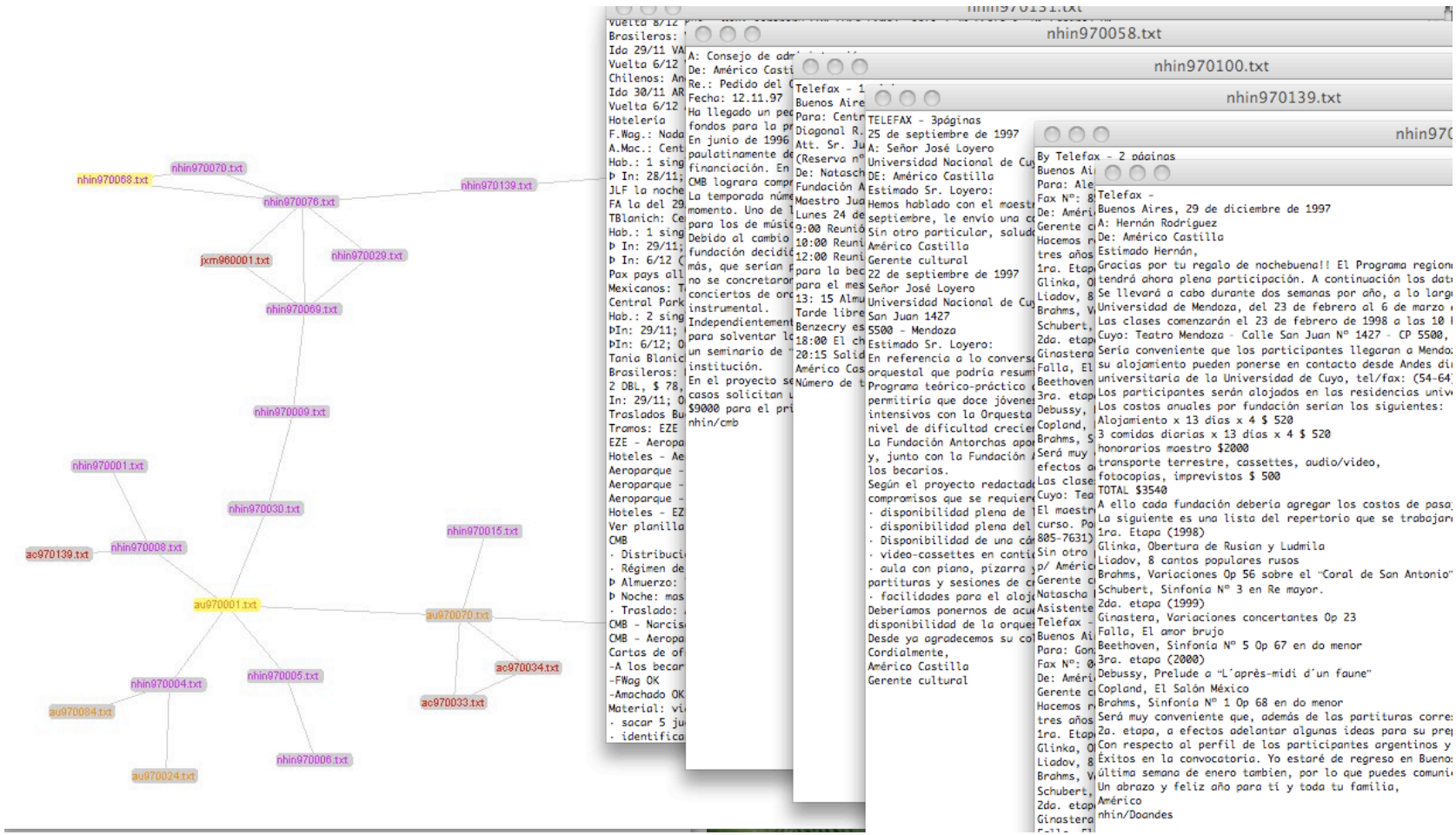




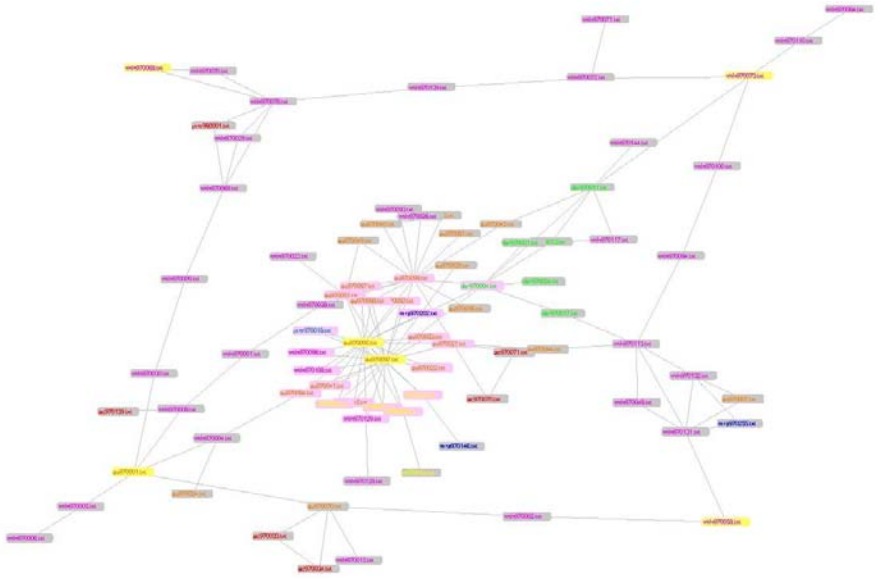
Compute relationships through bioinformatics-inspired method called “paragraph alignment”

- Unstructured collection of electronic records of different authors, topics, sizes with not much of an a-priori organization

Stories


nhin970070.txt
nhin970068.txt
nhin970076.txt
nhin970139.txt
jxm960001.txt
nhin970029.txt
nhin970069.txt
nhin970009.txt
nhin970001.txt
nhin970030.txt
nhin970016.txt
ac970139.txt
nhin970008.txt
au970001.txt
au970070.txt
ac970034.txt
nhin970004.txt
nhin970005.txt
ac970033.txt
au970084.txt
nhin970006.txt
au970024.txt
nhin970058.txt
nhin970100.txt
nhin970139.txt
TELEFAX - 3páginas
25 de septiembre de 1997
A: Señor José Loyero
Universidad Nacional de Cuyo
DE: Américo Castilla
Estimado Sr. Loyero:
Hemos hablado con el maestro
septiembre, le envío una co
Sin otro particular, saludo
Américo Castilla
Gerente cultural
22 de septiembre de 1997
Señor José Loyero
Universidad Nacional de Cuyo
San Juan 1427
5500 - Mendoza
Estimado Sr. Loyero:
En referencia a lo conversa
orquestal que podría resumi
Programa teórico-práctico
permitiría que doce jóvenes
intensivos con la Orquesta
nivel de dificultad crecier
La Fundación Antorchas apor
y, junto con la Fundación
los becarios.
Según el proyecto redactado
compromisos que se requiere
· disponibilidad plena de
· disponibilidad plena del
· Disponibilidad de una cá
· video-cassettes en canti
· aula con piano, pizarra y
partituras y sesiones de c
· facilidades para el aloj
Deberíamos ponernos de acue
disponibilidad de la orques
Desde ya agradecemos su co
Cordialmente,
Américo Castilla
Gerente cultural
Telefax -
Buenos Aires, 29 de diciembre de 1997
A: Hernán Rodríguez
De: Américo Castilla
Estimado Hernán,
Gracias por tu regalo de nochebuena!! El Programa regiona
tendrá ahora plena participación. A continuación los dat
Se llevará a cabo durante dos semanas por año, a lo larg
Universidad de Mendoza, del 23 de febrero al 6 de marzo
Las clases comenzarán el 23 de febrero de 1998 a las 10
Cuyo: Teatro Mendoza - Calle San Juan N° 1427 - CP 5500,
Sería conveniente que los participantes llegaran a Mendo
su alojamiento pueden ponerse en contacto desde Andes di
universitaria de la Universidad de Cuyo, tel/fax: (54-64
Los participantes serán alojados en las residencias univ
Los costos anuales por fundación serían los siguientes:
Alojamiento x 13 días x 4 $ 520
3 comidas diarias x 13 días x 4 $ 520
honorarios maestro $2000
transporte terrestre, cassettes, audio/video,
fotocopias, imprevistos $ 500
TOTAL $3540
A ello cada fundación debería agregar los costos de pasa
La siguiente es una lista del repertorio que se trabajar
1ra. Etapa (1998)
Glinka, Obertura de Rusian y Ludmila
Liadov, 8 cantos populares rusos
Brahms, Variaciones Op 56 sobre el "Coral de San Antonio"
Schubert, Sinfonía N° 3 en Re mayor.
2da. etapa (1999)
Ginastera, Variaciones concertantes Op 23
Falla, El amor brujo
Beethoven, Sinfonía N° 5 Op 67 en do menor
3ra. etapa (2000)
Debussy, Prelude a "L'après-midi d'un faune"
Copland, El Salón México
Brahms, Sinfonía N° 1 Op 68 en do menor
Será muy conveniente que, además de las partituras corre
2a. etapa, a efectos adelantar algunas ideas para su pre
Con respecto al perfil de los participantes argentinos y
Éxitos en la convocatoria. Yo estaré de regreso en Buenos
última semana de enero tambien, por lo que puedes comuni
Un abrazo y feliz año para ti y toda tu familia,
Américo
nhin/Doandes

Discovering context



- Explore and discover relationships between records and their authors
  - Context
    - Evidenced of cooperative writing
    - Work-processes

Conclusions

- Preliminary research
- Usability, display, interoperability and direct adaptation to archival tasks need to be resolved
  - EAD, JHOVE, PREMIS, METS
- Results stored in a database,
  - can be combined in myriad ways and with other tools to make abstractions, synthesis, and new discoveries
- Find useful/new visual metaphors