



This Ain't Your Father's Digital Data

Another Perspective on Legal Information
(<http://www ldc.upenn.edu/>)

Christopher Cieri

Linguistic Data Consortium
University of Pennsylvania
Philadelphia, PA 19104



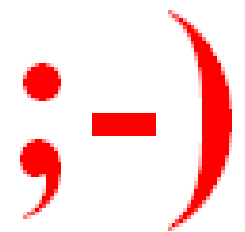
Introductory Questions

- What is the structure of the presentation?
 - Background: 10 minutes
 - Building Corpora of Digital Data: 45 minutes
- Who is this guy?
 - And how did he come up with such a wacky topic?
- What is this Linguistic something-or-other?
 - And why should I care?
- What are the Goals?
 - connect LDC and language technology developers to the community of researchers, educators and information managers in Law
 - keep in touch with databases and related efforts in Law
 - open discussions on how LDC data can be useful to this community
 - identify opportunities for collaboration



LDC

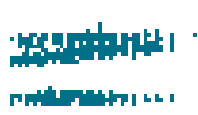
- **Linguistic Data Consortium**
 - A non-profit activity of the University of Pennsylvania,
 - an open consortium of Universities, Government agencies and Companies founded in 1992 with DARPA/NSF support
 - to collect, create, prepare and distribute language data
 - for education, research, clinical practice and technology development related to language
- **What is language data?**
 - Text, image, audio and video containing language.
 - Overlaps with many professions including:
 - » medicine: Disarthric Speech
 - » government: Resource Management
 - » communications: Topic Detection & Tracking
 - » law: Juris, Hansard, United Nations Parallel Texts





Corpora in Law

- **Hansard**
 - proceedings of Canadian parliament in French and English
 - mid 1970's to 1988
 - 3300 files (645MBs compressed)
- **JURIS**
 - from Department of Justice Information Retrieval System, all data undisputably in the public domain
 - Administrative Law, Briefs, Case Law, Executive Orders, Federal Regulations, International Agreements, Regulations, Statutory Law, Tax Law from 1700 to 1990
 - 650,000 documents in 1650 files (960MBs compressed)
- **UN Parallel Text**
 - from UN electronic text archives in English, French, Spanish
 - 1988-1993
 - 92,000 files, 2.5GBs text



U.N. Example

LETTER DATED 24 DECEMBER 1992 FROM THE PERMANENT REPRESENTATIVE OF IRAQ TO THE UNITED NATIONS ADDRESSED TO THE SECRETARY-GENERAL

93_00004.ENG: On instructions from my Government, I wish to inform you of fresh Iranian violations of the cease-fire between the two countries during the period from 17 November to 20 December 1992.

93_00004.FRE: D'ordre de mon gouvernement, j'ai l'honneur de vous informer des violations des dispositions du cessez-le-feu entre l'Iraq et l'Iran commises par l'Iran au cours de la période allant du 17 novembre au 20 décembre 1992.

93_00004.SPA: Cumpliendo instrucciones de mi Gobierno, deseo informarle de las nuevas violaciones de la cesación del fuego entre ambos países cometidas por el Irán entre el 17 de noviembre y el 20 de diciembre de 1992.



Model

- Originally, distribute data created by researchers
- Increasingly, create databases in response to community needs
- Reach out to new communities
- Make data available to all
 - Members join yearly and receive all corpora published in that year for free.
 - Some corpora can also be sold to non-members.
- Act as mediator of intellectual property creating usage agreements that satisfy data owners and speed acquisition of data

Why?

- Research and development often require huge amounts of data
 - thousands of hours of speech
 - tens of millions of words of text
- Corpus building on a large scale requires:
 - specialized equipment
 - specially trained staff
- Effort required to establish ad hoc collection
 - beyond reach of most non-profits
 - typically not cost-effective for corporations
 - unpalatable to government agencies who want to see public monies used efficiently
- Stable sources of reference data promote discussion, comparison & evaluation

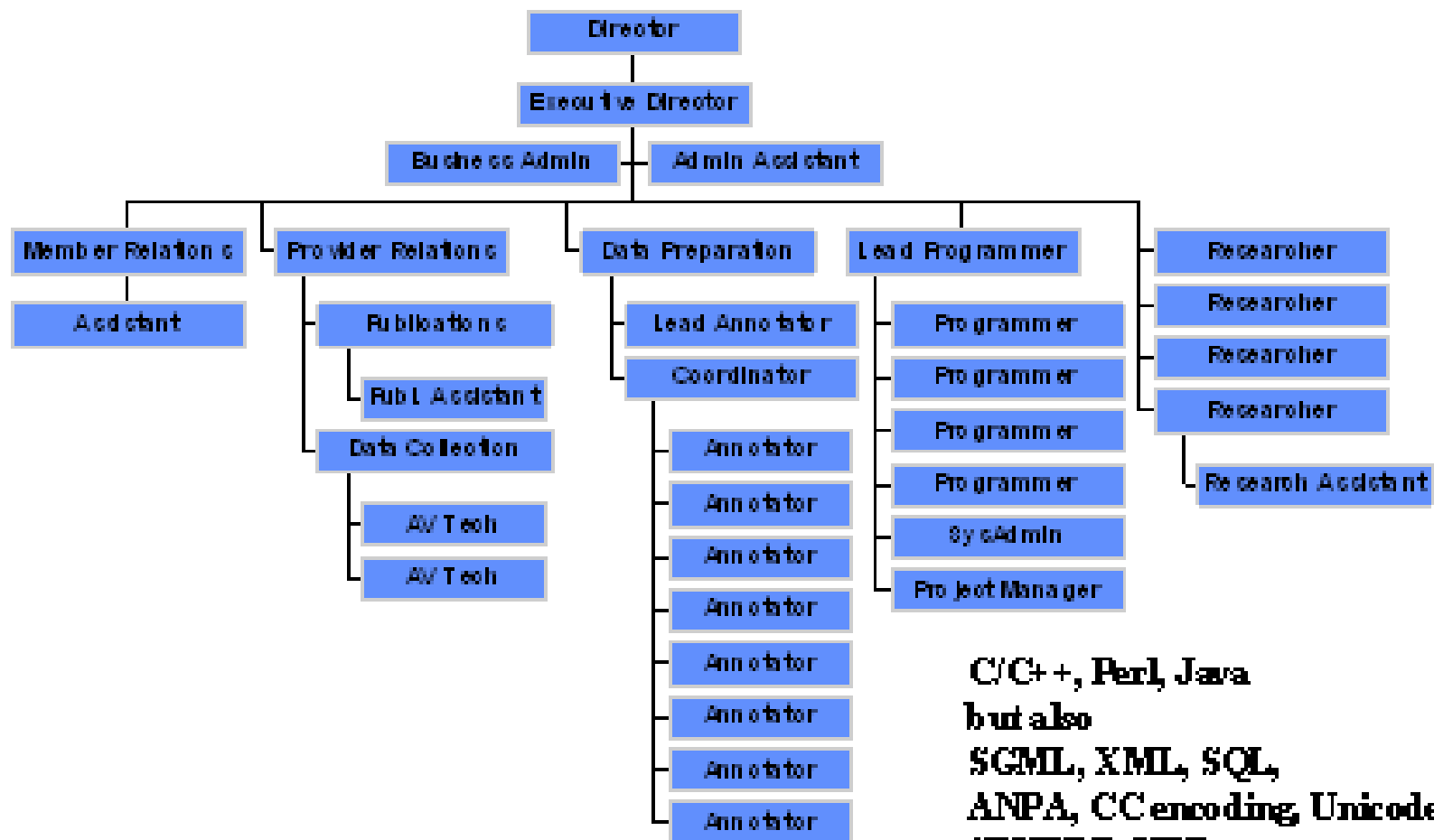


Specialized Equipment

- **Infrastructure**
 - 3 Sun E4000 multi-processors, >1GB RAM each
 - 1.2TB RAID disk shared
 - 1 TB near-line storage
 - 3.5TB tape robot for backup
 - Administrative Server: NT, 27GB RAID, tape robot
- **Special**
 - Dedicated fiber-optic network
 - Satellite Downlink - multifunction, receives VOA
 - Telephone Collection - T1, 24 lines, 54GB RAID
 - Annotation Workstations
 - » > 60 Sparcs, >20 PCs (few Macs for compatibility)
 - Miscellaneous Collection Hardware
 - » AV receivers & recorders, CC decoders, DATs, etc.



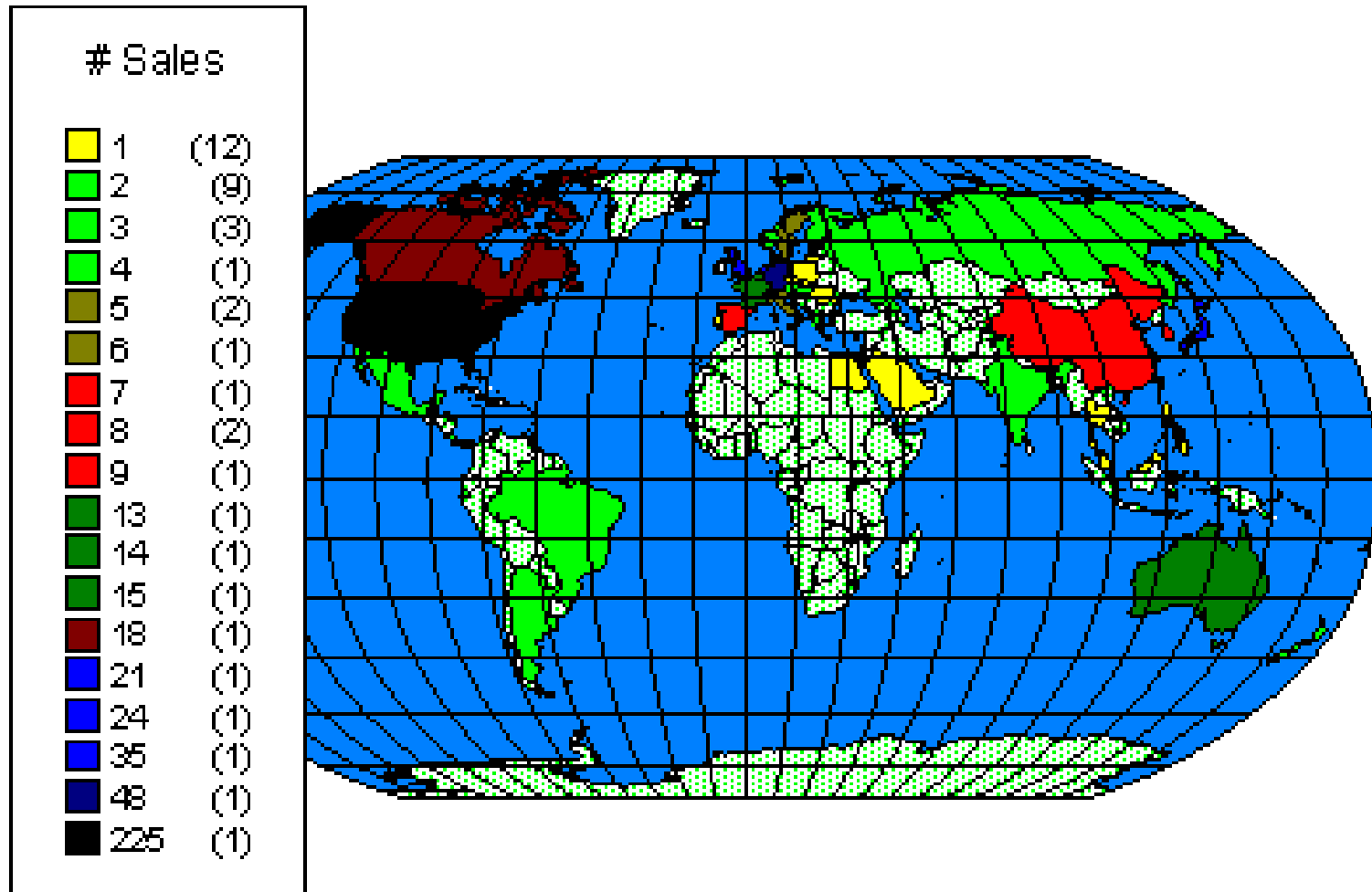
Specially Trained Staff



C/C++, Perl, Java
but also
SGML, XML, SQL,
ANPA, CC encoding, Unicode,
SPHERE, UTF
PerlTK, PerlLib, Emacs-Lisp



Data Available to All



■ CALI 99 - June 17-19, 1999 - Eugene, Oregon



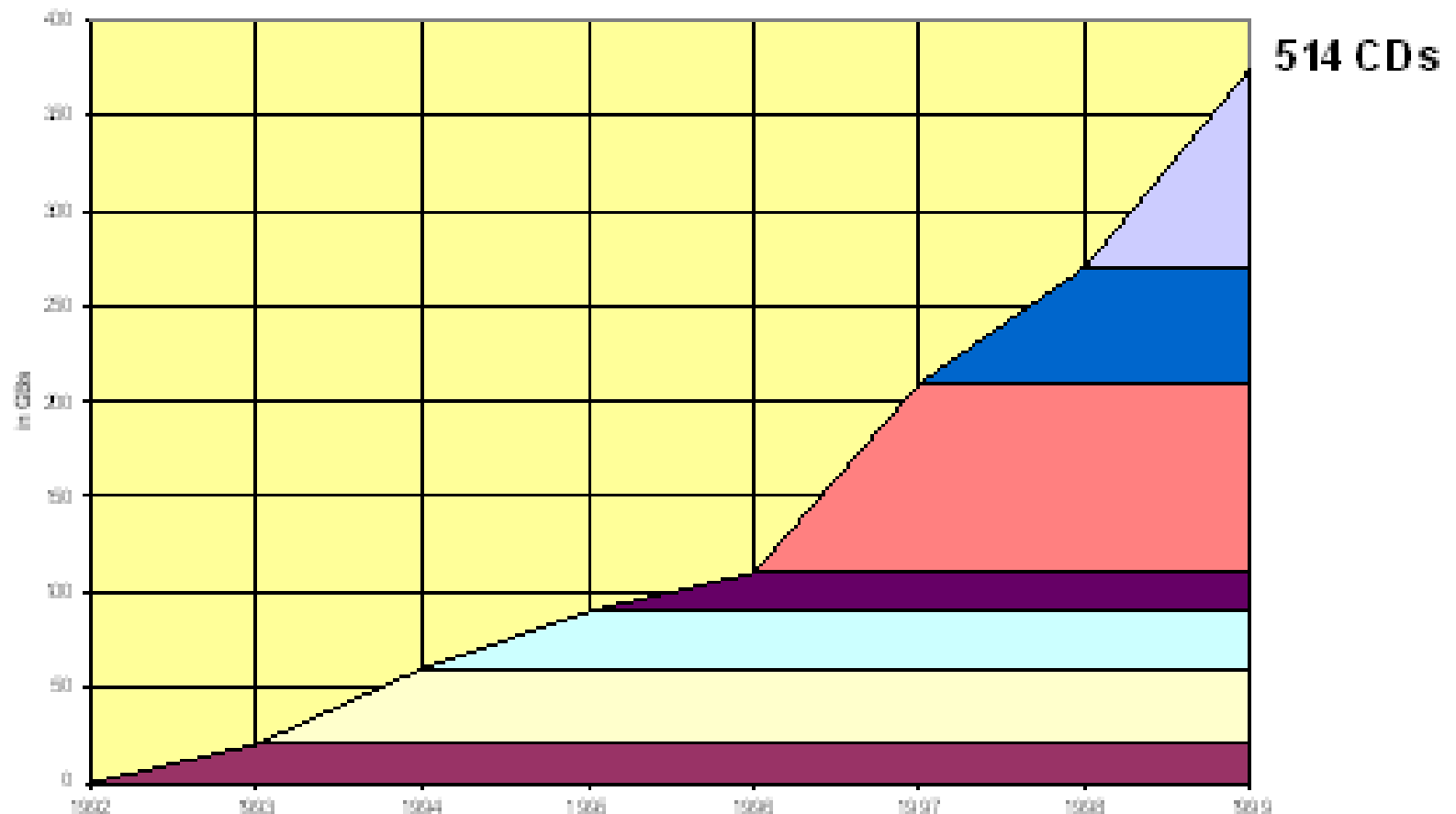
Library
Development
Center

Library
Development
Center

Library
Development
Center

How Much Data?

Data Growth



- **LDC Experience**
 - over 150 corpora published
 - over 650 organizations served
 - » Engineering, Language, Regional Studies, Medicine and Law
 - nearly 10,000 distributions transacted
- **Is it time yet?**
 - National Gallery of the Spoken Word - project proposed by the Vincent Voice Library and backed by NSF
 - SignComp - first ever digital repository of and collection facility for video data in support of gestural and sign language research
- **Let's Talk**



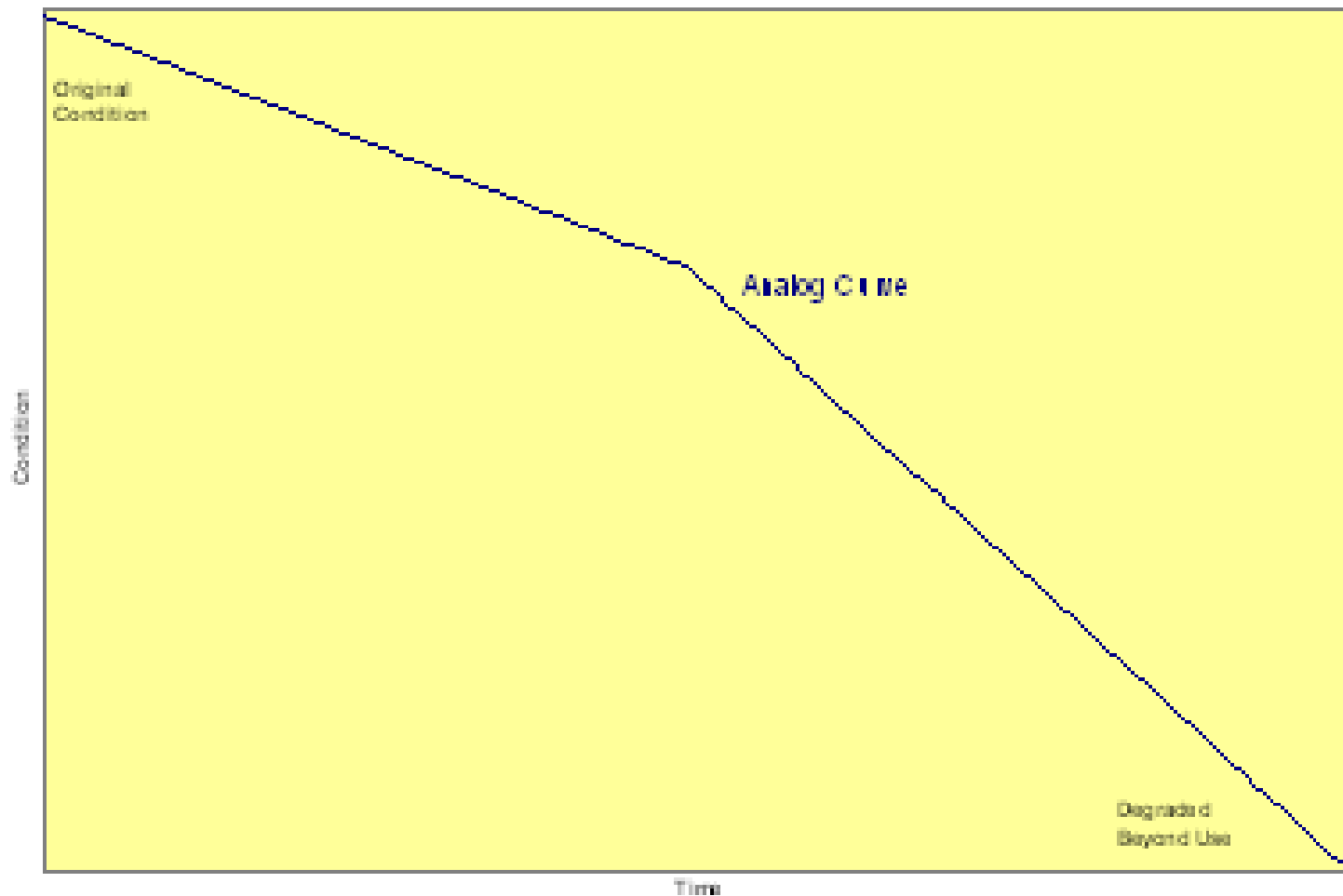
Why Digital?

- Offer another mode of presentation to accommodate new learning/working styles
- Can be dynamic; data and code can interact
- Accessible faster & more easily even from a distance
 - faculty & staff at conferences, distance learning
- More easily searchable
- Redundant - multiple simultaneous users
- Use can be more easily monitored, profiled
- Deflects demand from non-digital resources
- And otherwise aids in preservation
- In CALI, it offers an alternative to embedding data into code.

- Structured demographic data for Admissions, Placement; patron & collection data for Library
- Full Text for Legal Research
- Virtual Video Tour
- Talking Resumes
- Video examples of interviews, negotiations
- Data to go!
 - Faculty summer vacations
- Analysis of performance at Moot Court competition
- Woodhouse's Virtual Supreme Court and Family Law Class

- **Speech Recognition systems to recognize legal terms and adjust to legal language**
 - phoneme models and language models
 - commercial approaches (cf Gregory Clinton's paper)
 - unsupervised approaches
- **Speech Synthesis systems to provide improved access to visually challenged**
- **Machine Translation of foreign government documents**
 - new initiatives underway to revive machine translation
- **Message Understanding**
 - extracts structured data from prose (www.nist.gov)

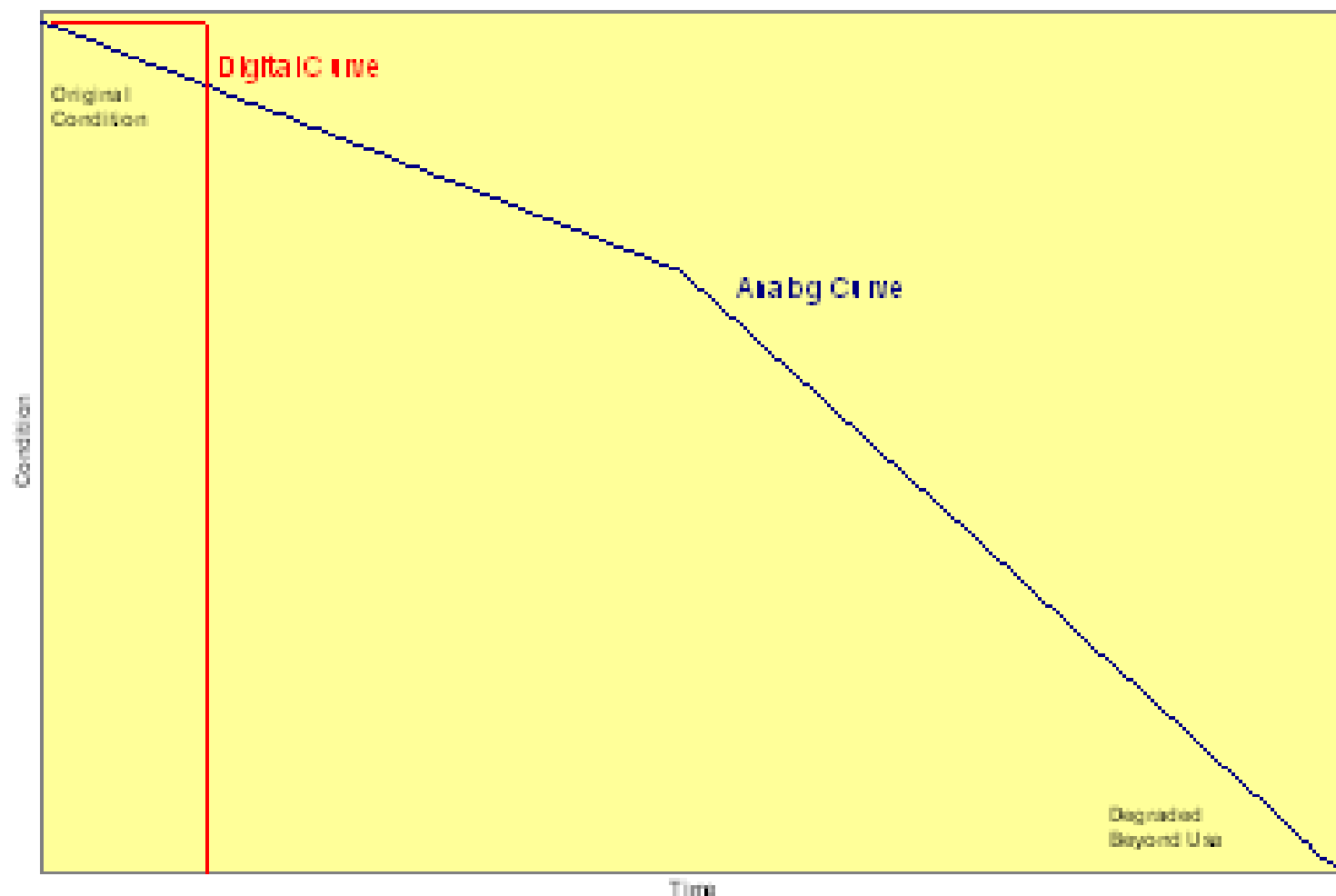
Degradation Curves





Preservation

Degradation Curves





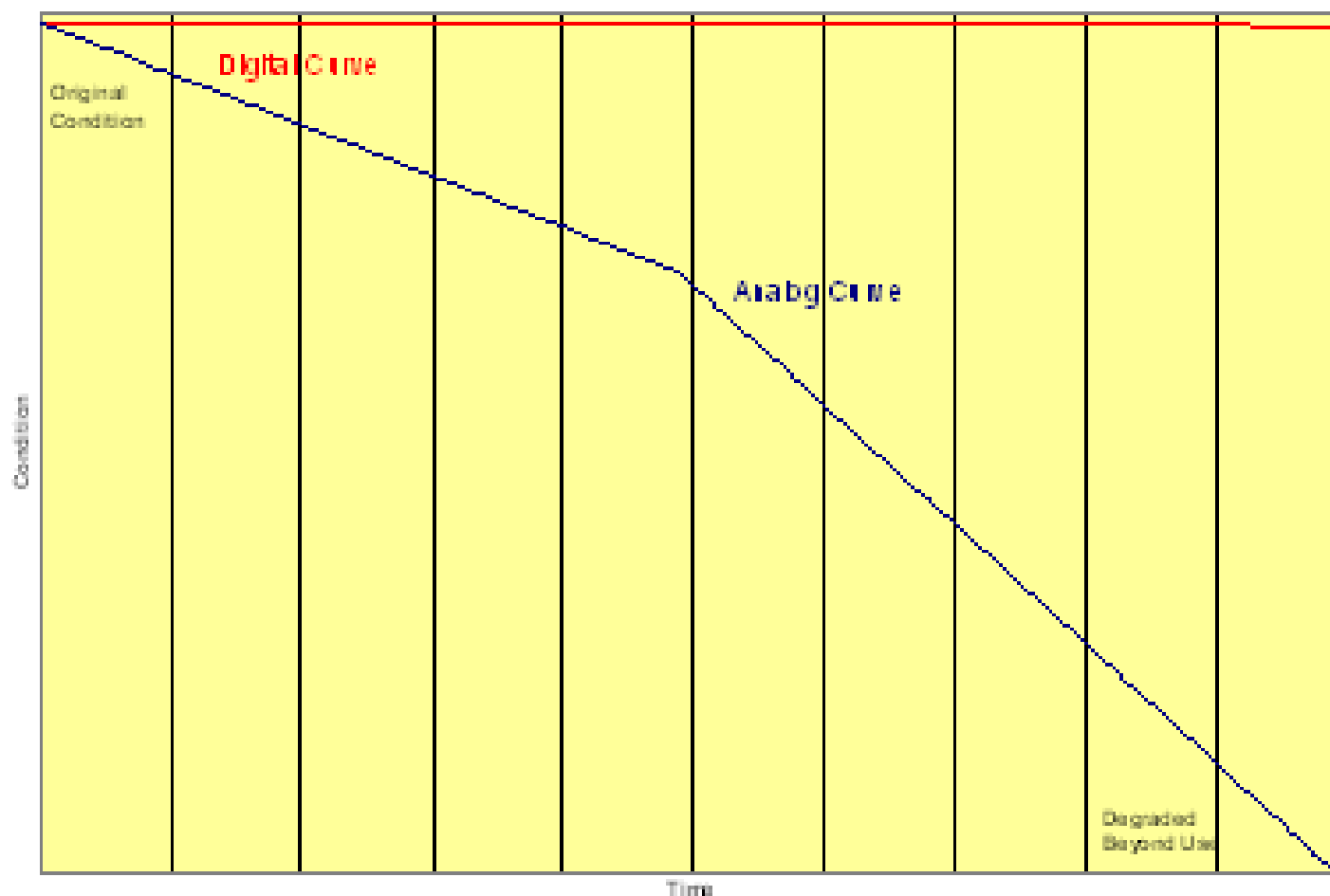
University of
Digital Collections

University of
Digital Collections

University of
Digital Collections

Preservation

Degradation Curves





Types

- Digital Data was traditionally structured data.
- Law has focussed on full text.
- Imaging is certainly coming into its own.
 - multiple Digital Library projects
 - current storage reaching necessary levels
 - » 160,000 pages on my laptop
- Technology for audio is prime.
- Video will follow shortly



Creating Digital Data

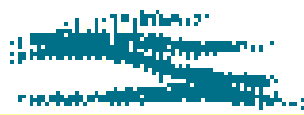
- Planning
- Acquisition of Originals
- Assessment
- Collection
- Segmentation
- Annotation (adding meta-data)
- Quality Assurance
- Preparation for Distribution
- Distribution

- What is the purpose of the project?
- How will the data be distributed?
 - Conflicts develop here because network bandwidth can lag behind local bandwidth.
- Originals
 - Nature of content
 - Physical Medium
 - » dimensions, physical attributes, robustness, quality, prospects for “cleaning”
- Cost Options & Staff Requirements
 - for internal or outsourced projects
- Process & Timelines
 - which tasks can be parallel which must be serial



Collection

- Sources
 - **Not just paper!**
 - Other databases on the WWW, typesetters files, broadcast, interviews
- Resolution - degree to which an analog signal is sampled to produce a digital artefact
 - 2-600 dpi graphics, 11,16,22,44KHz audio, 30-60fps video
- Quantization - range of values any single sample can have
 - 2 byte text, bitonal, gray-scale or color graphics at various depths, 16 bit audio
- This area developing rapidly. If much time passes between initial planning and kick-off, another planning cycle is necessary.



Collection Model

Limits of Biological System



Collection Model

Limits of Biological System

Full Information Capture



Collection Model

Limits of Biological System

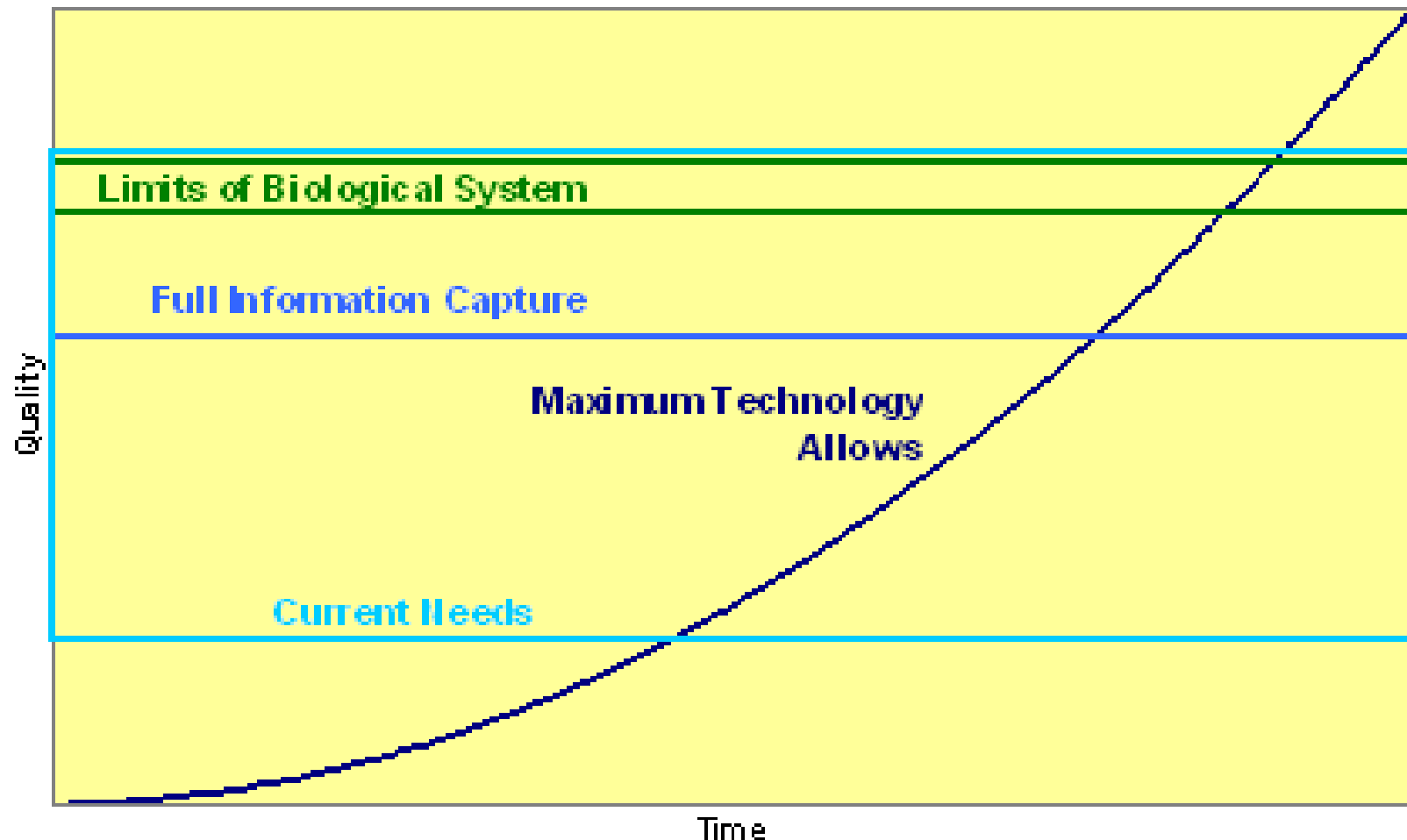
Full Information Capture

Current Needs



Collection Model

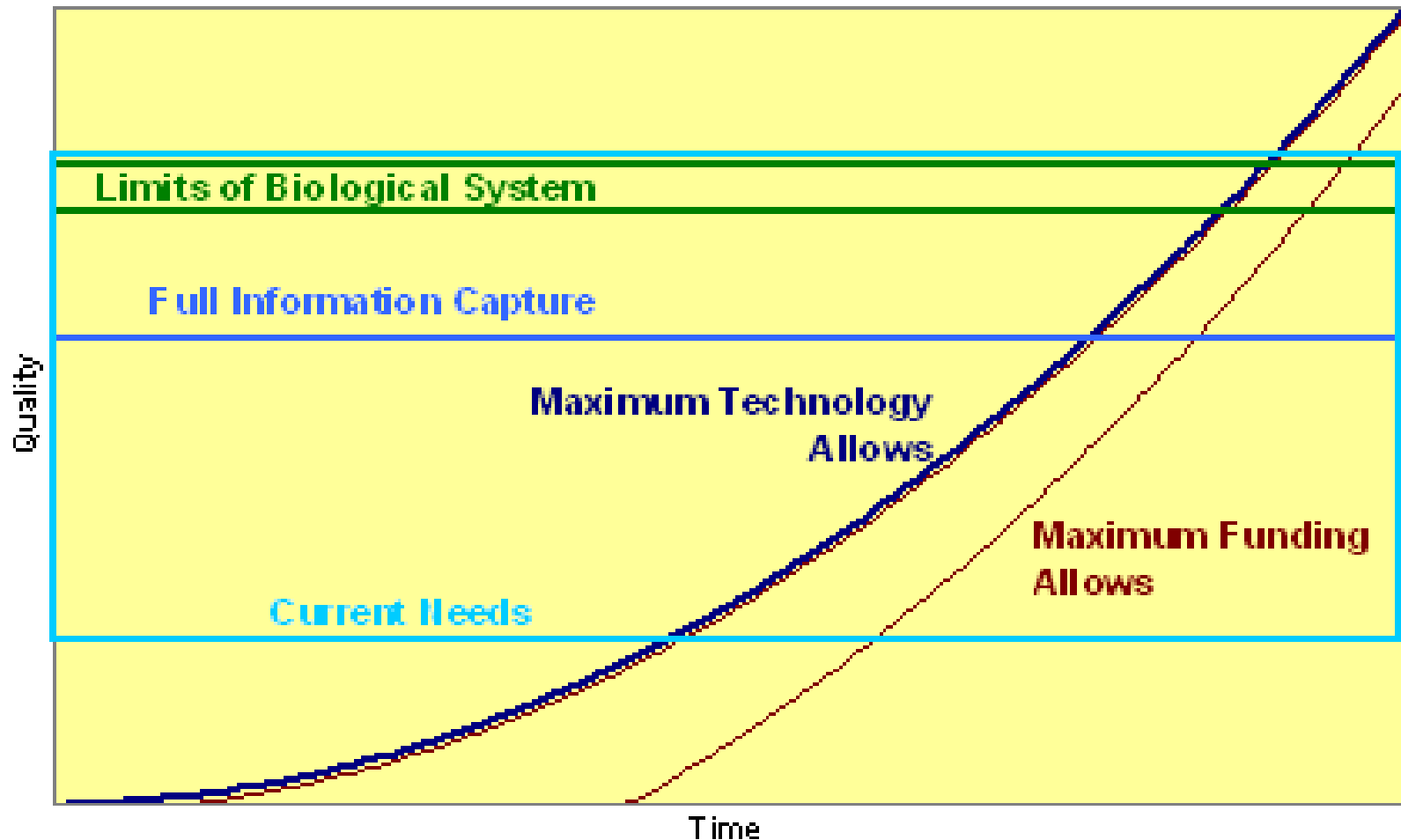
Options for Setting Quality





Collection Model

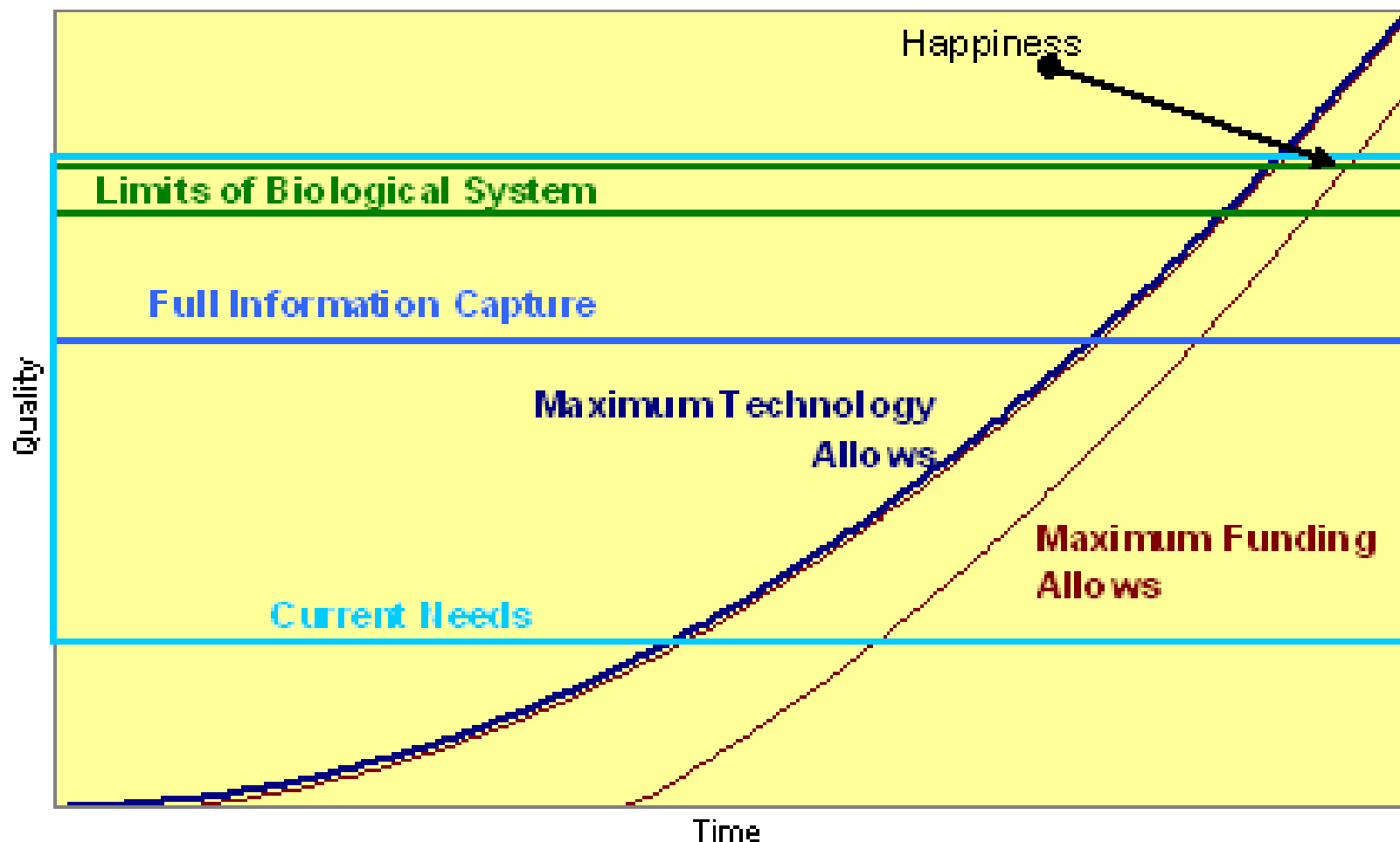
Options for Setting Quality





Collection Model

Options for Setting Quality

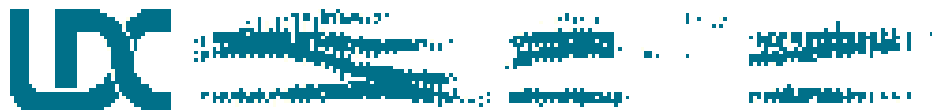




Segmentation

- Actually or virtually dividing a data glob into component parts.
 - Records and fields in a structured database
 - Chapters and sections of a book
 - Speaker turns, stories or sections in an audio or video collection.
- Granularity will depend upon intended use.
- And will change over time
- Stand-off approach is more suitable to change than actually chopping files

Original Audio												
Attorney												
Client												
Good												
Bad												



Annotation

- Any process that adds value to raw data.
- Date & authority fields on structured data
- Categorization of texts (ie. Cataloguing)
- Identification of sections of text
- Adding/identifying meta-data
- Transcribing speech
- Identifying participants
- Highlighting important points



Quality Assurance

- Terms with an example from Cataloguing
- Precision
 - attempt to find incorrect assignments of an annotation
 - 100%
- Recall
 - attempt to find failed assignments of an annotation
 - 10-20%
- Discrepancy
 - resolve disagreements among annotators
 - 100%
- Measures of Inter-annotator Agreement
 - measure and analyze sources of disagreement
 - 5-10%

Distribution Issues

- Proactive or Reactive digitization
 - either needs a reactive component
- Intellectual property/confidentiality issues
 - What limits do they impose?
- WWW based distribution and bandwidth
- Will data be available in another medium
- Preprocess multiple formats or create on the fly?
- LDC Online Model
 - searching by topic in the Topic Detection and Tracking case
 - other examples: www ldc upenn edu, Select LDC Online

- Digital data is more than just text and images.
- Corpora in Law support not only research and education in Law but also technology development and evaluation for the legal market.
- Standards and Infrastructure are ripe to support digital data in all media - expensively still in video.
- There is still some need for end-user authoring tools.