The Draft NAACCR XML Data Exchange Standard: Keeping Pace with the New Era in Cancer Surveillance and Research

Presented by:
Eric B. Durbin, DrPH, MS
Assistant Professor, Division of Biomedical Informatics
Director of Cancer Informatics, Kentucky Cancer Registry
Director, Cancer Research Informatics Shared Resource Facility
Markey Cancer Center
University of Kentucky

Keeping Pace....

Keeping Pace with the New Era of Cancer Research

- Surveillance data must support a new paradigm in population-based cancer research
  - Genomics
  - Biorepositories
  - Virtual tissue repositories
  - Virtual pooled registries

Keeping Pace with the New Era in Cancer Surveillance

- Surveillance must accommodate a new paradigm in the electronic health arena
- Increased reliance on electronic data sources and automation
  - Rapidly emerging and changing
    - Electronic pathology
    - Electronic health records
    - Precision medicine
  - BIG data
    - Large data elements such as Next Generation Sequencing (NGS)

Is the current NAACCR fixed-width data exchange standard keeping pace with the new era?

1. Yes
2. No
3. Not Sure

Significance of the NAACCR Data Exchange Standard

- One of the most important technological advances in cancer surveillance
- Ensures syntactic and semantic interoperability
- Universally adopted by all cancer registries in North America
  - Hospital cancer registries
  - Regional cancer registries
  - Central cancer registries
  - National cancer surveillance agencies

Continued NAACCR Ownership of Our Data Exchange Standard Essential

- Necessary to meet our cancer surveillance mission
- NAACCR membership best understands our needs, data sources, operations, limited resources and data uses
- Cancer surveillance data are special and unique
  - Highly structured hierarchical data curated from a variety of reporting sources over time
  - Patient -> Cancer Diagnosis -> Treatment -> Outcomes
- Not just a collection of clinical events and observations

NAACCR Volume II Fixed-width Format

- NAACCR record divided into well defined sections
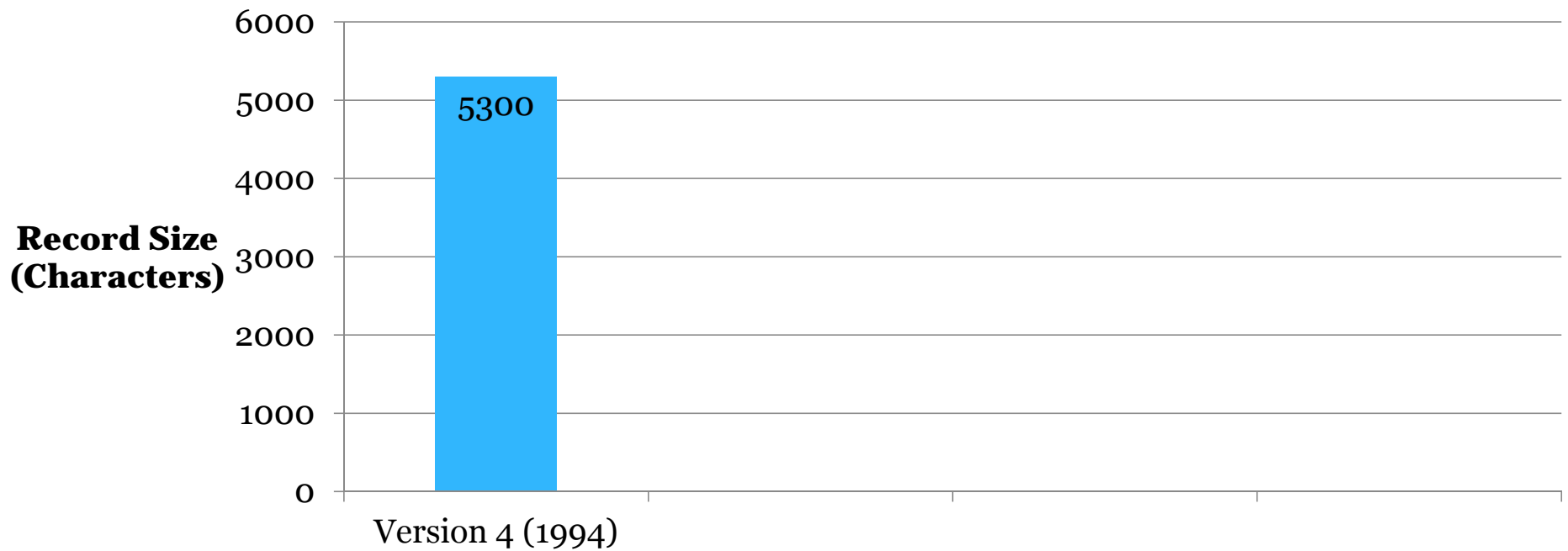- Data elements occupy a fixed location defined by the column position

Section	Column Start	Column End
Record Identification	1	94
Patient Demographics	95	527
Cancer Identification	528	690
Hospital-Specific	691	903
Stage/Prognostic Factors	904	1435
Treatment (First and Subsequent)	1436	1887
Edit Overrides/Conversion History/Admin	1888	2115
Follow-up/Recurrence/Death	2116	2339
Special Use (State Requestor)	2340	3339
Confidential (Patient/Other/Pathology)	3340	5564
Text	5565	22824

Advantages of a Fixed-width Format

- Relative ease of use
- Parsing tools easy to develop
- Surveillance community is familiar with this standard
- Impact of changes is fairly predictable
  - Major overhaul expected every 7 years or so
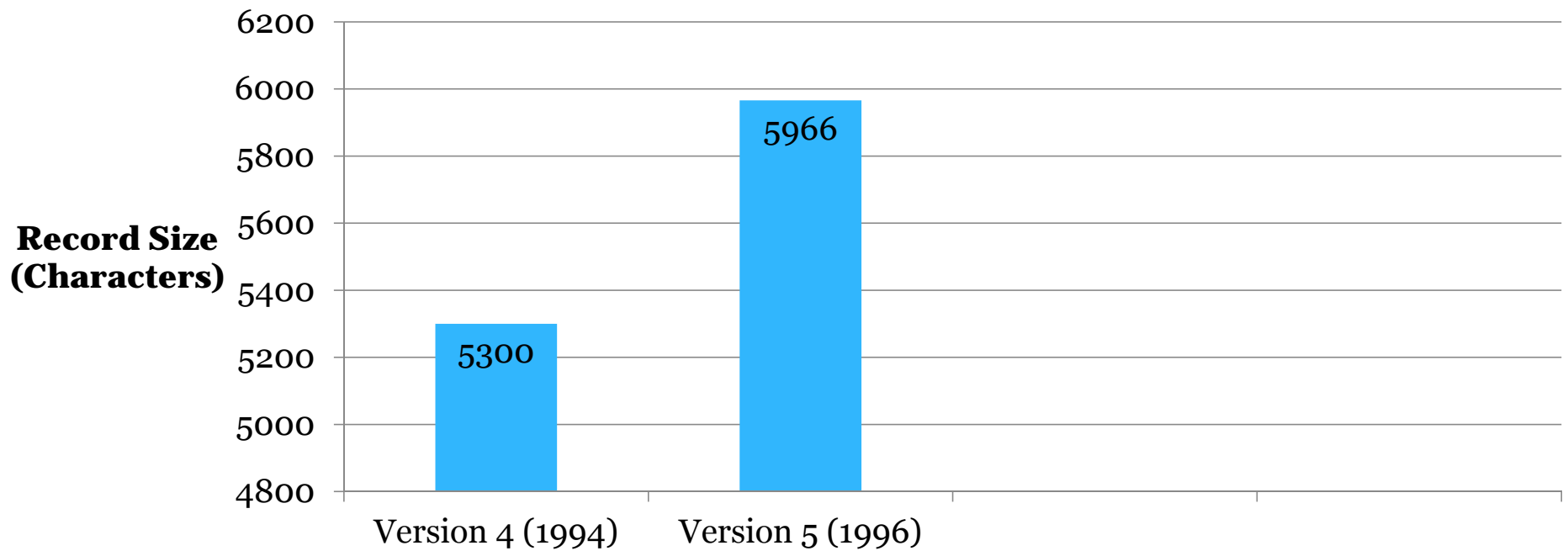- Does it evolve well with growing needs?

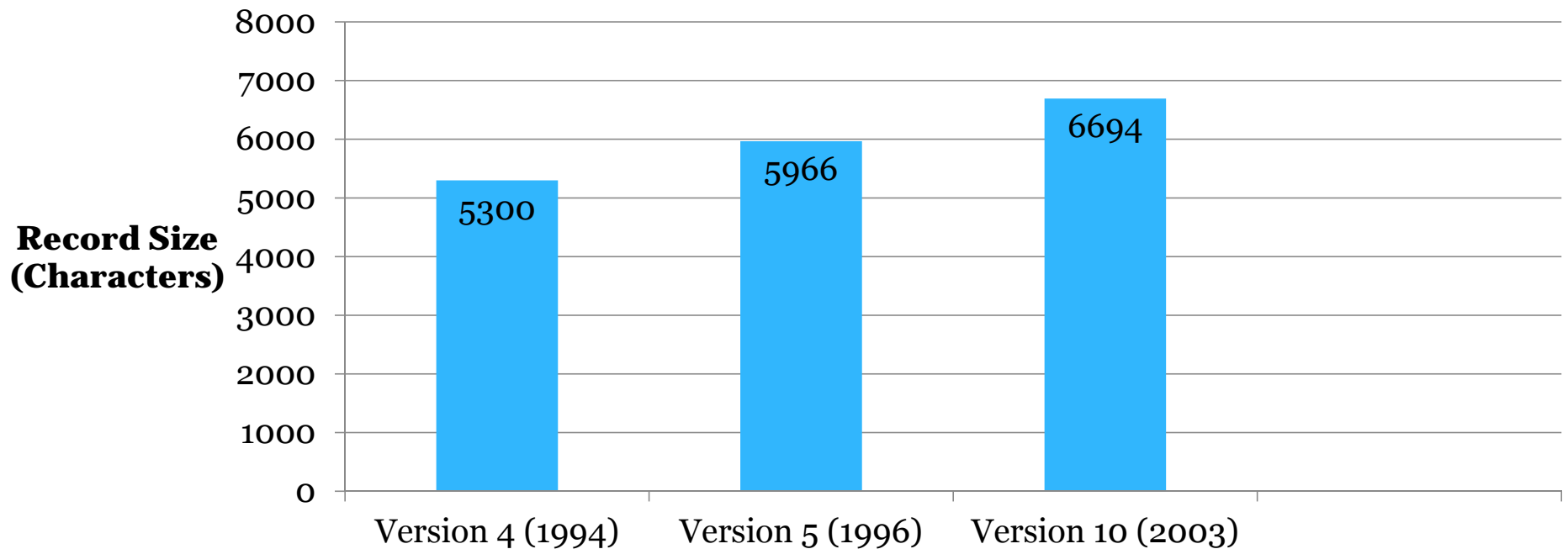# NAACCR Volume II Evolution Over Time

Major Revisions

Record Size (Characters)

A bar chart showing record size in characters. The Y-axis ranges from 0 to 6000 in increments of 1000. A single bar for Version 4 (1994) reaches 5300.

Version	Record Size
Version 4 (1994)	5300

NAACCR Volume II Evolution Over Time

Major Revisions

Record Size (Characters)

Version	Record Size
Version 4 (1994)	5300
Version 5 (1996)	5966

NAACCR Volume II Evolution Over Time

Major Revisions

Record Size (Characters)

8000	
7000	
6000	
5000	
4000	
3000	
2000	
1000	
0	

- Version 4 (1994): 5300
- Version 5 (1996): 5966
- Version 10 (2003): 6694

NAACCR Volume II Evolution Over Time

Major Revisions

Record Size (Characters)

- Version 4 (1994): 5300
- Version 5 (1996): 5966
- Version 10 (2003): 6694
- Version 12 (2010): 22824
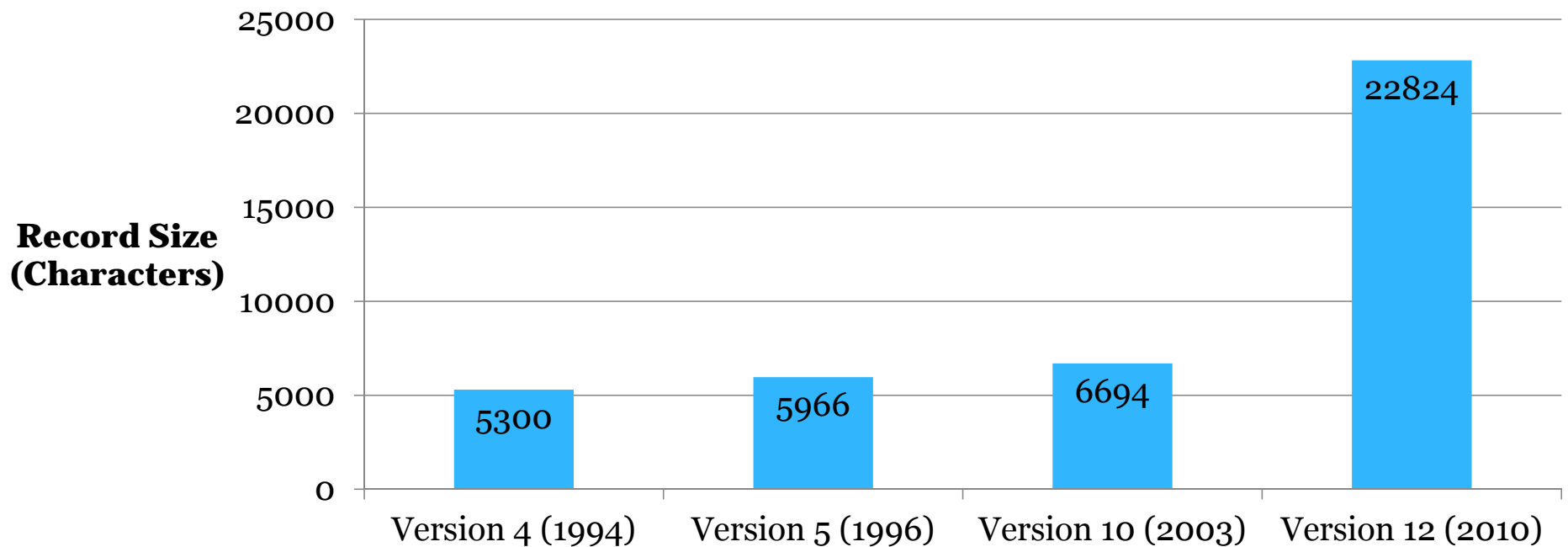
NAACCR Volume II Evolution Over Time
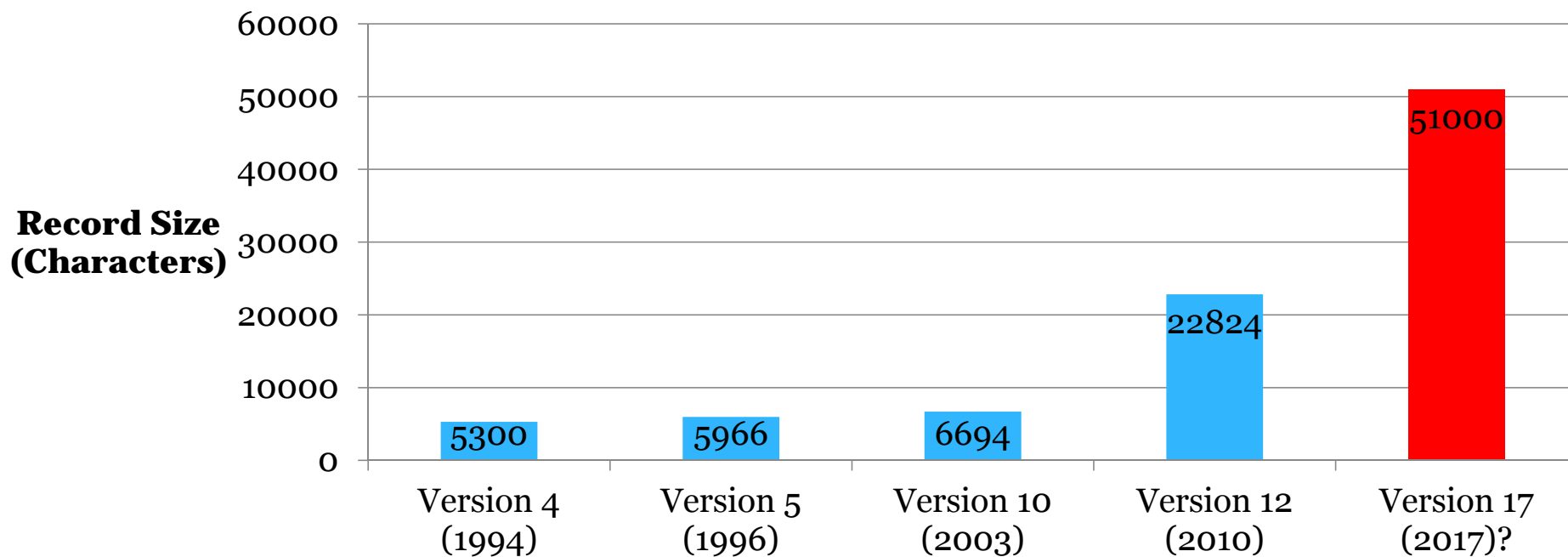
Major Revisions

Challenges of the NAACCR Fixed-width Layout

- Doesn't scale well over time
- Major revisions break all existing software implementations
- Previous NAACCR data files rendered incompatible with each new major revision

Challenges of a Fixed-width Data Exchange Standard

- Limits what can be transmitted
  - If it doesn't fit, throw it away
  - No longer reflects what is captured in cancer surveillance databases
  - Limited treatment
  - Limited text
- Not very flexible
  - If it doesn't fit, forget about it
  - Can't accommodate sizeable or complex additional data elements
- Not particularly interoperable with other standards
  - No metadata to identify context or content

What should we do to keep pace?

1. Nothing (deny we have a problem)
2. Wait for other initiatives to render cancer surveillance obsolete
3. Take early retirement
4. Adopt a more modern data transmission standard

NAACCR XML: A Potential Solution to Enable Cancer Surveillance to Keep Pace

- NAACCR Committees and Task Forces
  - Studied problem since 2006
  - Evaluated other standards
    - CSV, HL7 2.x, HL7 CDA
- Current NAACCR XML Task Force
  - Formed in 2014
  - Chaired by Isaac Hands
  - Participation by national standard setters (ACoS, CDC, SEER, NAACCR), software vendors, central registries
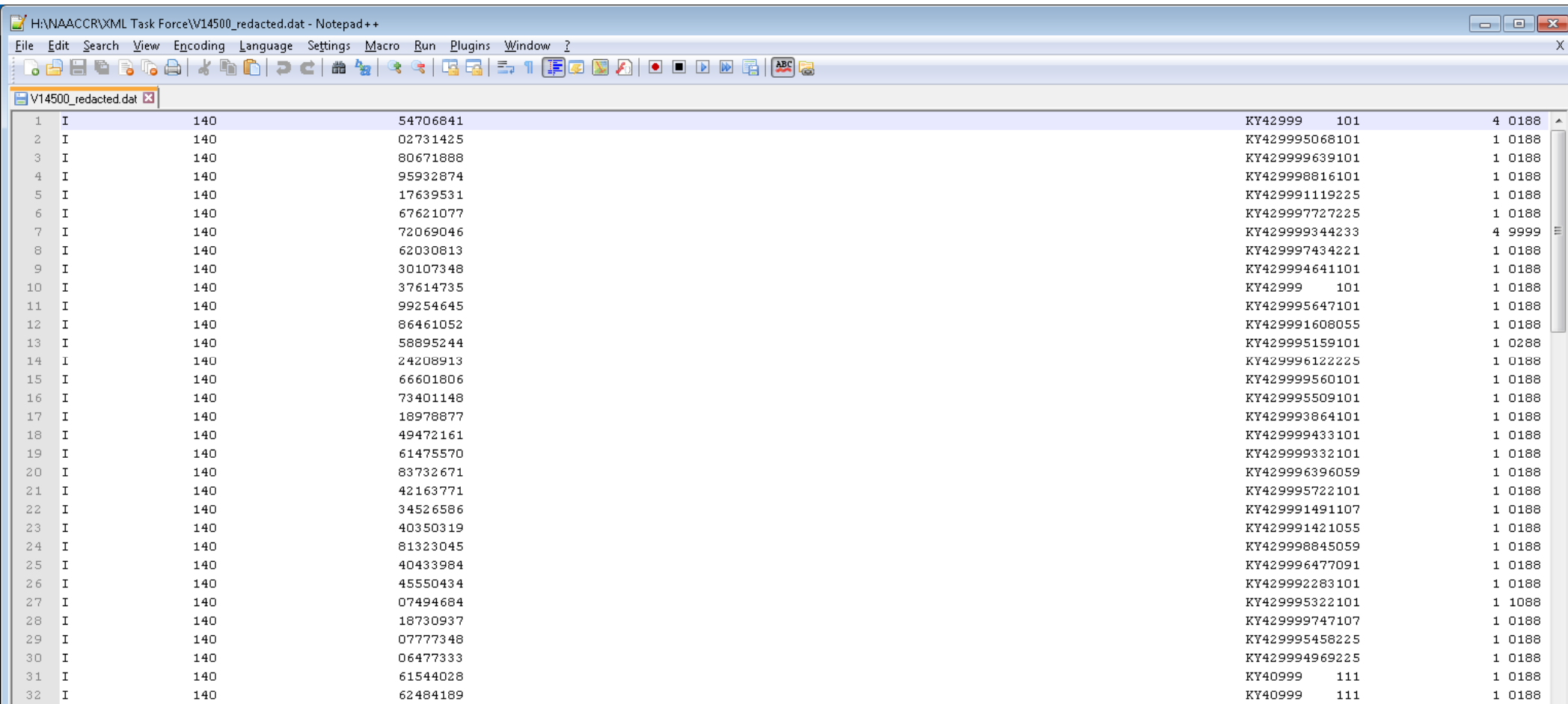
Fundamental XML Task Force Objectives

- Do no harm
- Develop an alternative draft standard in XML
- Develop, test and evaluate XML for common use cases
  - Hospital to central registry
  - Central registry to national organization
  - Registry to researcher
- Ensure an upgrade path for all registries
  - Allow gradual adoption
  - Low cost
  - Sustainable

NAACCR XML: A Simple Solution

- Rather than define NAACCR elements by column position in a fixed-width record
- Define NAACCR elements with XML tags
  - Informational tags that specify record version
  - Informational tags that specify record sections
  - Informational tags that specify individual data items

From this:

To this:

Advantages of the NAACCR XML Standard

- Eliminates loss of data
  - Certain elements can be nested and repeated
  - Certain elements no longer constrained by a fixed number of characters
- Introduces greater flexibility
  - New elements can be introduced without breaking previous software
  - Unknown elements can be ignored if received
- Introduces greater interoperability
  - XML tags communicate context (NAACCR, version, record type)
  - XML tags communicate meaning

NAACCR XML Supports a Gradual Adoption Path for Registries

- Permits continued use of fixed-width standard for as long as necessary
- XML conversion software tool has been developed and released
  - Developed by Information Management Services, Inc. (IMS)
    - Fabian Depry
  - Free and Open Source
  - Downloadable today

http://naaccrxml.org

- Documentation
  - Frequently Asked Questions
  - Implementation Guide
- XML Conversion Tool
  - XML -> Flat (fixed-width)
  - Flat -> XML

NAACCR XML Concurrent Session 1: Section B
Today at 1:30PM,
Grand Ballroom B – Level 2

- Moderator
  - Gary Levin, Florida Cancer Data System
- Introduction to XML: How it Works and What it Offers Us
  - Rich Pinder, Los Angeles Cancer Surveillance Program
- Creating, Transmitting, and Working with a NAACCR XML file
  - Isaac Hands, Kentucky Cancer Registry
- XML Tools - Examples to Slice, Dice, Load, and Analyze XML Data
  - Fabian Depry, Information Management Services, Inc. (SEER)
- Discussion

See NAACCR XML in Action! Exhibit Hall

- During poster sessions AND
- 7:30am on Wednesday and Thursday

- Live demonstration of the new software tool
  - Convert NAACCR flat file to NAACCR XML
  - Convert NAACCR XML to NAACCR flat file
- Live demonstration of EDITS on a converted NAACR XML file
- Live demonstration of SAS analysis on a NAACCR XML file
- Other amazing demonstrations
- See it, test drive it, ask questions

Do you plan to attend the XML breakout or visit NAACCR XML in Action demonstrations?

1. Yes
2. No
3. Hello early retirement!

Acknowledgements: XML Task Force Members

- Isaac Hands, Chair, Kentucky Cancer Registry

- Chris Johnson, Idaho Cancer Registry
- Dave Stinchcomb, Westat
- Dustin Dennison, NAACCR
- Eric Durbin, Kentucky Cancer Registry
- Fabian Depry, IMS/SEER
- Gary Levin, Florida Cancer Data System
- Gemma Lee, Cancer Care Ontario
- Heidi Gianella, ELEKTA
- Joe Rogers, CDC/NPCR

- Ken Pristas, ACoS/NCDB
- Kim Best, Rocky Mountain Cancer Data Systems
- Lori Havener, NAACCR
- Mark Hunzeker, Rocky Mountain Cancer Data Systems
- Rich Pinder, Los Angeles Cancer Surveillance System
- Ryan McCabe, ACoS/NCDB
- Roger Friedman, Northrop Grumman, CDC/NPCR

Not Keeping Pace...

Tale of Verve

Questions/Discussion