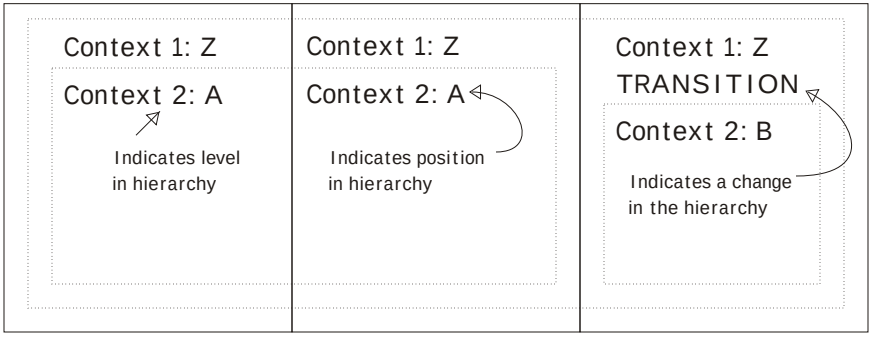
WHAT DO I MEAN?

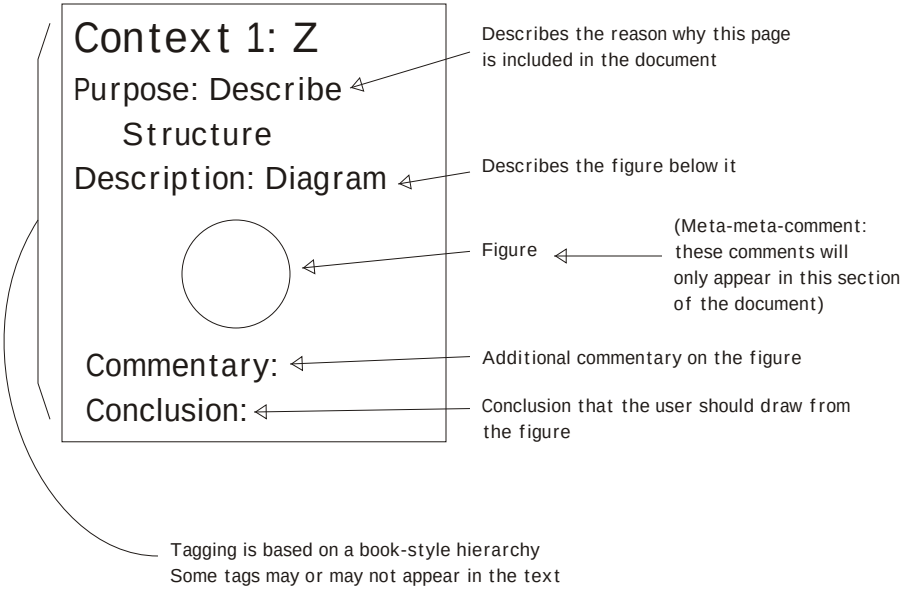
Insights and Issues Concerning Human Readable Data Formats

Ming-Yee Iu
#96 010 411
December 14, 2001

Context 1: Human-Readable Data

METADATA





Context 1: Human-Readable Data
Context 2: Introduction

Purpose: Anecdote showing importance of human-readable data
Description: My old high school spreadsheet from Economics class created in an old spreadsheet program I no longer own

ÿV‘J4HøÌ2,äAphÊ6¼B šÿ–"·…}3 %·Œ6Zí\ÿ ÿH "›¢BÏMÂXR Ó•¥$à sVµ©© Ðo<gŸX‰Ïÿ¬ §.þâY• LÜg ÿÿ:}ÿ±ÐÇÿ ÿâ ßš=V• ŸÆC°¯q‹ 8¢ÉÑSÎÿšeFÊÛÿMÿÈéjLÞàÿ˜ÿßÀ, éì¥õöÍÿÐÿ±¬Zc‘®"Le• N?PppÐÿÿ ÿçÿiiiwww†††——• • • ¤¤¤²²²ËËË ×××ÝÝÝãããêêêñññøøø²Áf€¿x Æðð²¤ÿÿ³ÿÑŽ£ÃÜ7 žTv®pxžÁƒd¿ ¤ƒÓÑ?2ÿ}

Conclusion: The document in unusable and the data will be lost

Context 1: Human-Readable Data
Context 2: Introduction

Purpose: As technology develops, programs become obsolete and old data becomes inaccessible

Description: Development of different Microsoft Word file formats

1989	*Word 1.0*
1990	*Word 1.1*
1991	*Word 2.0*
1993	*Word 6.0*
1995	*Word 95/7.0*
1996	*Word 97*
1999	*Word 2000*

Conclusion: Document formats change quickly

Context 1: Human-Readable Data
Context 2: Introduction

Purpose: The idea that a universal format will solve data obsolescence is not valid

Description: Creations dates of different character sets

1960s EBCDIC
1965/67 ASCII
1972 ISO-646
1980s ISO-8859
1990/91 ISO-10646/Unicode

Conclusion: Even standards that seem stable actually need to change fairly often to adapt to new developments

Context 1: Human-Readable Data
Context 2: Introduction

Purpose: The longevity of English suggests that human-readable data can extend the useful lives of data
Description: History of the English language

7th-11th century	*Old English*
12th-15th century	*Middle English*
16th-21st century	*Modern English*

Commentary: Though English does change, the tools and knowledge needed to decode older versions of English are well-documented
Conclusion: Data encoded in English can potentially last a long time

Context 1: Human-Readable Data
Context 2: Introduction

Purpose: Where can developers turn to for guidance on how to develop these formats?
Description: Related fields to human-readable data

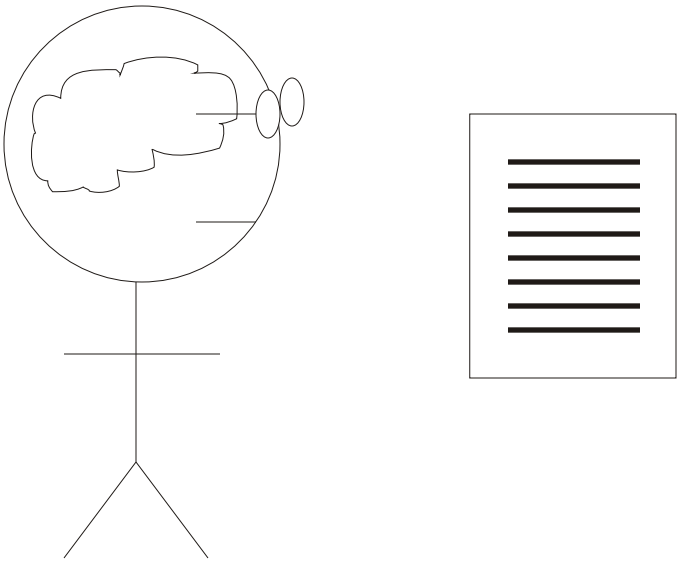
Anti-cryptography
(Communication with alien races)
Archivists and library preservationists
Archaeologists
Psychologists
Graphic Designers

Commentary: Though many fields are related to human-readable data, little directly relevant research has been undertaken
Conclusion: Can these fields provide any insight into the problem?

Context 1: Human-Readable Data
TRANSITION from Context 2: Introduction
Context 2: The Nature of the Problem

Purpose: What is the fundamental problem in creating and interpreting human-readable data?

Description: To interact with something, a user must bring a body of knowledge to the task



Conclusion: File formats must encode this knowledge describing the task or be based on some universal understanding that does not need to encoded

Context 1: Human-Readable Data
Context 2: The Nature of the Problem

Purpose: What knowledge needs to be encoded in a file and what does not?
Description: Shared context as time passes

After 1 year
- *Is encoding the same?*
- *Is user background the same?*
- *Is culture the same?*
- *Is language the same?*
- *Is humanity the same?*

After 10 years?

After 100 years?

Conclusion: We can only assume that humanity and language remain the same over time

Context 1: Human-Readable Data
Context 2: The Nature of the Problem
TRANSITION to New Level of Context
Context 3: Useful things to Encode

Purpose: The information that archivists have found useful to encode in documents
Description: Raw data

AKFJGURESSG
DFDGJA@#%
A132jasdf8jASDJ2fdfgkadfg;
DF38aksieygrgkj34pfs8bfj34
oerda;kei4e9843oigmkvfdpi
3498df;fgskl34p98s;jsfoi;roie
r89sjjmv'43]sf;fg;sfdg

Conclusion: Obviously, the data of the document itself needs to be stored in a document

Context 1: Human-Readable Data
Context 2: The Nature of the Problem
Context 3: Useful things to Encode

Description: Data dictionary (describe encoding of raw data)

1. *1.32*
2. *5.36*
3. *10.15*

Conclusion: One must know how to read the data too, so the data dictionary must also be encoded

Context 1: Human-Readable Data
Context 2: The Nature of the Problem
Context 3: Useful things to Encode

Description: Code book (describes meaning of variables used in document)

Rotten Heads: $1.32
Small Heads: $5.36
Large Ripe Heads: $10.15

Conclusion: This info must be encoded as meta-data in the file, but how?

Context 1: Human-Readable Data
Context 2: The Nature of the Problem
Context 3: Useful things to Encode

Description: Documentation of context and research methods

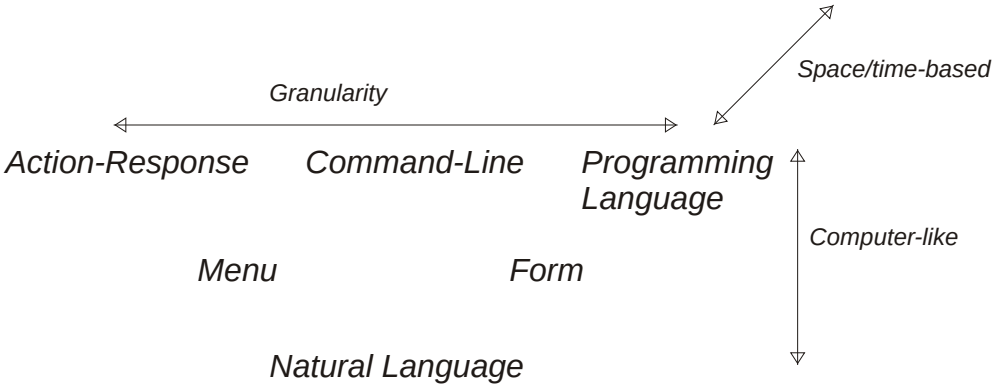
Lettuce Dealer Price List

Rotten Heads: $1.32
Small Heads: $5.36
Large Ripe Heads: $10.15

Conclusion: This data also needs to be encoded as meta-data in the file, but how?

Context 1: Human-Readable Data
TRANSITION from Context 2: The Nature of the Problem
Using Connection: Even if we know what to encode, what insight do we have about how to encode it?
Context 2: Breakdown by Axes

Purpose: HCI researchers gain insight into UIs by breaking problems into different axes
Description: Axes of different dialog styles

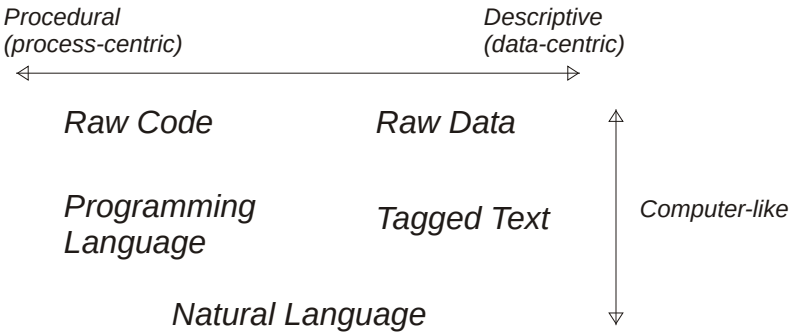


Conclusion: Can a similar approach be used with human-readable data?

Context 1: Human-Readable Data
Context 2: Breakdown by Axes

Purpose: One possible axes breakdown of human-readable data encoding types

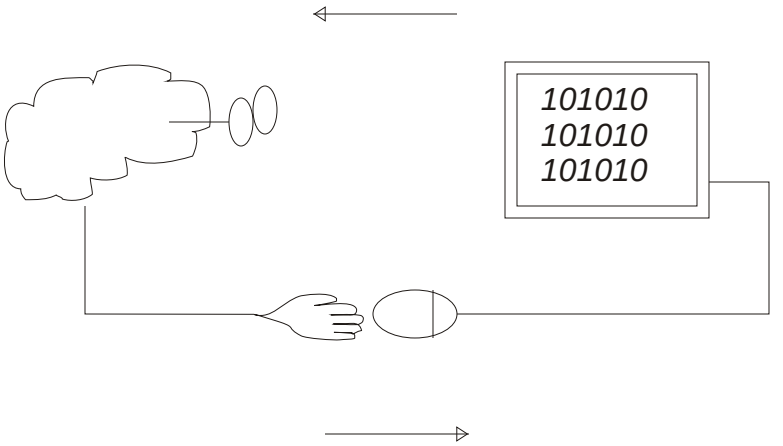
Description: Cross-section of human-readable data



Conclusion: Such a breakdown of the encoding problem yields few insights

Context 1: Human-Readable Data
TRANSITION from Context 2: Breakdown by Axes
Using Connection: Breakdown by axes provided little insight, but will another approach help break down the problem into easier to study pieces?
Context 2: Breakdown Using Interaction Models

Purpose: HCI researchers can reduce a UI problem by modeling its interactions
Description: Diagram of elements of interaction between user and computer

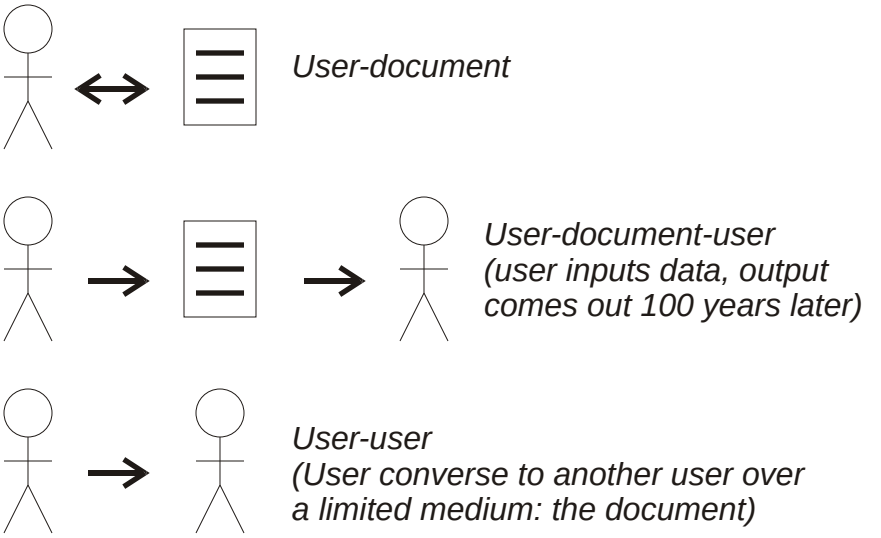


Conclusion: Examining only the interaction of a user with a computer helps focus research on a specific aspect of the problem

Context 1: Human-Readable Data
Context 2: Breakdown Using Interaction Models

Purpose: Although documents generally do not accept input, looking at documents from a different perspective may allow us to apply the model

Description: Different interactions involved of users viewing human-readable data

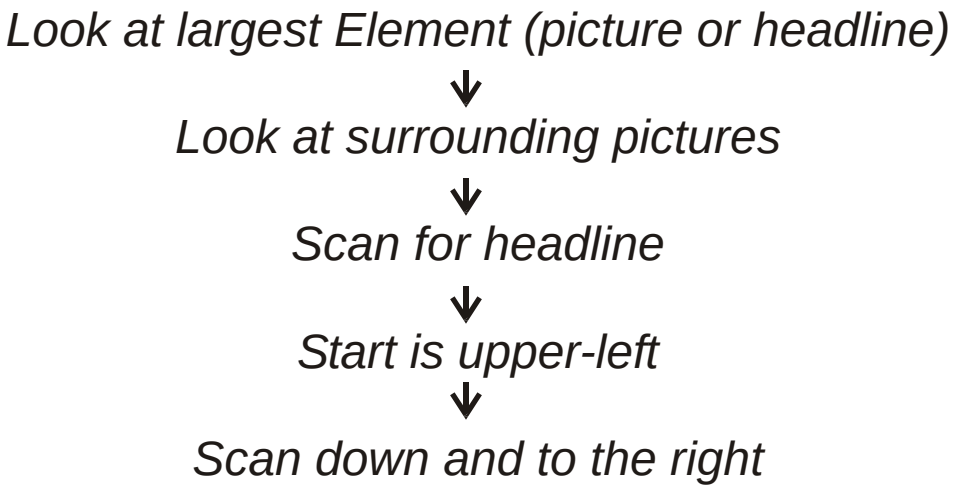


Conclusion: Lack of feedback in documents limit the usefulness of applying interaction models to human-readable data

Context 1: Human-Readable Data
Context 2: Breakdown Using Interaction Models

Purpose: Though graphic design and typography have limited feedback, they too can use interaction models

Description: State machine of a user looking at a graphic layout

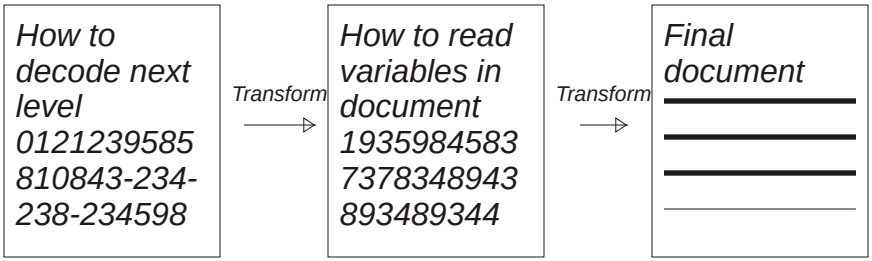


Commentary: In this interaction, the layout of the document controls the user and pushes the user through transitions into various states

Conclusion: Years of psychological research and anecdotal evidence were required to develop this sort of knowledge. It is still premature to do this for human-readable data

Context 1: Human-Readable Data
Context 2: Breakdown Using Interaction Models

Purpose: Alternately, interaction can be designed into the file format
Description: In each state, document reveals some information and describes the transformation needed to move to the next state



Conclusion: This design may help avoid information overload and provide a learning gradient to direct the user

Context 1: Human-Readable Data
TRANSITION from Context 2:
Breakdown Using Interaction Models
Using Connection: Are there other possible ways of breaking down the problem?
Context 2: Taxonomy

Purpose: The purpose of taxonomies

Description: Colin Wheildon's experiment on the merits of different layout styles

	Comprehension Level		
	Good	*Fair*	*Poor*
Layout with serif body type	*67*	*19*	*14*
Layout with sans-serif body type	*12*	*23*	*65*

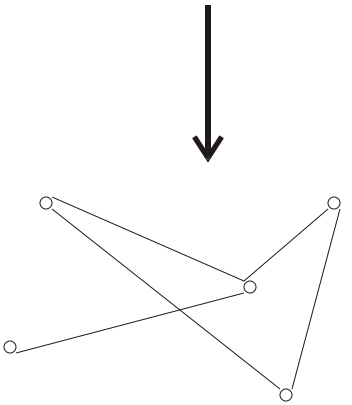
Commentary: Developing a taxonomy requires us to examine different human-readable formats, but unfortunately, no repository of this sort of information exists

Conclusion: A taxonomy allows for the discovery of new solutions to a problem and allows for experiments accessing the value of these different solutions

Context 1: Human-Readable Data
Context 2: Taxonomy

Purpose: How can a taxonomy be built?
Description: Information can be represented with propositional logic, which can be represented as a graph

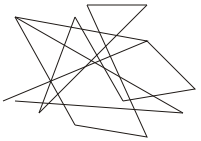
IsAtWaterloo(me) = true
For all x, IsAtWaterloo(x) implies IsBitter(x)



Conclusion: All data can be represented as graphs

Context 1: Human-Readable Data
Context 2: Taxonomy

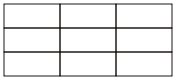
Purpose: What are the dominant encodings for human-readable data
Description: Different graph types and their text encodings



Arbitrary graph

```
A  B  1
C  D  2
D  A  3
```

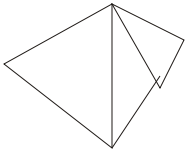
Relational



Grid

```
AASD   BSDSD   C
DSD    ESSD    F
GSDS   HSD     I
```

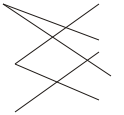
Table



Planar

```
        Asdfasd
Asdfasd
            Asdfasd
Asdfasd
                Asdfasd
```

Free form



Bipartite

```
Ads = adsf
Dsf = asdf
Eie = 234
```

Key-value pairs

Tree

```
(Sdfs
  (Adf
    (Df
      Df
    )
  )
)
```

Hierarchical

Commentary: Each encoding type can be augmented to give them a limited ability to encode the other graph types
Conclusion: By examining different graph types, we may be able to build a taxonomy of different structures for human-readable data

Context 1: Human-Readable Data
TRANSITION from Context 2: Taxonomy
Using Connection: Now that we know about different types of encodings for human-readable data, can we evaluate these encodings without performing experiments?
Context 2: General Insights

Purpose: Implicit knowledge of users
Description: Assumptions that users will make when examining data

Order from Chaos

-If a user recognizes a word or structure, that word or structure actually exists and is not a random artifact of the encoding

Metaconversations

-Metadata is distinctive from data
-Meta data is not part of the data

Consistency

-Format of entire file is similar
-Structure of file will not change arbitrarily

Conclusion: The design of human-readable data should not contradict implicit knowledge

Context 1: Human-Readable Data
Context 2: General Insights

Purpose: Two principles that can be applied to human-readable data formats

Grouping

Linear-Time

Context 1: Human-Readable Data
Context 2: General Insights

Purpose: Grouping principle
Description: Discussion of grouping

BECAUSE

People's brains are designed to observe things in parallel and process things sequentially thereby allowing it to observe groupings withing extensive amounts of processing

WE CAN

Imply relations between parts of data by grouping data together by proximity in space or in attributes

Commentary: Is there a benefit to adding additional grouping attributes to text such as <red>wall</red> <yellow>door</yellow> even though these attributes are not attributes of the actual data?
Conclusion: Grouping is an effective way to structure data

Context 1: Human-Readable Data
Context 2: General Insights

Purpose: Linear time principle as applied to hierarchies
Description: Discussion of linear time

BECAUSE

Text is read linearly in time making constructions such as triply center embedded sentences difficult to understand (e.g. “The audience the lecture I was attending was boring was asleep”)

WE CAN

Deduce that a hierarchical or more complicated structure is not effective

Commentary: Hierarchical structures do allow users to suppress unnecessary detail, but this is irrelevant in human-readable data
Commentary: Shallow hierarchical structures may still be ok
Commentary: Although code can be arbitrarily nested, programmers tend to prefer code that is divided into smaller groupings in a shallow hierarchy
Conclusion: Hierarchical and other more complicated structures may be ineffective

Context 1: Human-Readable Data
Context 2: General Insights

Purpose: Linear time principle as applied to metadata
Description: Discussion of linear time

BECAUSE

Text is read linearly in time meaning it is cumbersome to read an entire document

WE CAN

Deduce that describing the structure of a document at the beginning of the document is more useful than interspersing the information throughout the document

Conclusion: Instead of describing tags and structures where they occur in the document, it is more useful to describe them at the beginning of the document

Context 1: Human-Readable Data
TRANSITION from Context 2: General Insights
Context 2: Specific Insights

Purpose: Insights from dialogues
Description: Desirable properties of dialogues

Reference: cues refer to stuff we know

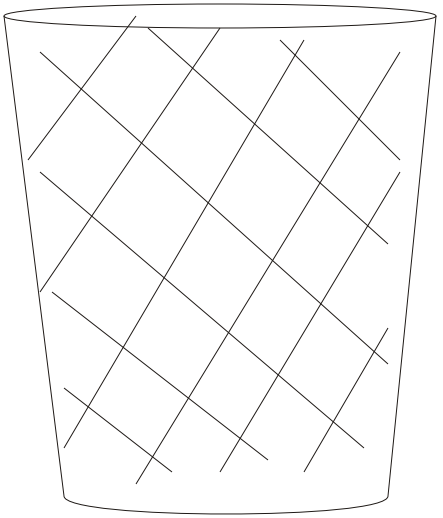
Consistency: no mix-up of conventions

Congruency: take advantage of collections of associations that users already know

Conclusion: Human-readable data should also have these properties

Context 1: Human-Readable Data
Context 2: Specific Insights

Purpose: Insights from icons
Description: Icons act as symbols suggesting functionality



Trash

Commentary: Icons are based on recognition whereas tags are based on a combination of recognition and problem-solving
Conclusion: Tags should be chosen so that they suggest their functionality

Context 1: Human-Readable Data
Context 2: Specific Insights

Purpose: Insights from command languages
Description: Choices inherent in command languages

Choice of commands
- *Random: short commands*
- *Parts: Rearrange parts and prefixes to get new commands*
- *Natural language: verbose*

Choices for attaching objects
- *Linguistic naturalness*
- *Consistent concept*
- *Prepositions (superior)*

Commentary: Commands are based on recall whereas tags are based on a combination of recognition and problem-solving
Conclusion: Building tags using common parts and with prepositions to denote parameters is effective

Context 1: Human-Readable Data
Context 2: Specific Insights

Purpose: Insights from forms
Description: Some comments on forms

Geometry and colour are important

Beginner users use descriptions of boxes to learn how to use the form

Experienced users are able to ignore the descriptions and fill in the boxes quickly

Conclusion: Natural language comments (descriptive text that is not part of the meaning of the document) can help users interpret documents without hindering automated parsing of the document

Context 1: Human-Readable Data
Context 2: Specific Insights

Purpose: Insights from UIs that deal with limited user memory
Description: To help users orient and navigate linear UIs, various memory aids are useful

History of how current location was reached

Landmarks showing current location in the hierarchy (e.g. Chapter 1.4.2)

Conclusion: Landmarks and history are useful features

Context 1: Human-Readable Data
TRANSITION from Context 2: Specific Insights
Context 2: Conclusion / A Case Study

Purpose: Introduce XML
Description: A sample XML document

```
<?xml version="1.0"?>

<!DOCTYPE html PUBLIC
"-//W3C//DTD XHTML 1.1 plus MathML 2.0//EN"
"http://www.w3.org/TR/MathML2/dtd/xhtml-math11-f.dtd">

<Html xmlns="http://www.w3.org/1999/xhtml">
   <body>
   <p>
        <math xmlns="http://www.w3.org/1998/Math/MathML">
                    <mtext> Theorem 1: </mtext>
        </math>
   </p>
   </body>
</html>
```

Conclusion: XML is a popular human-readable format and should be examined

Context 1: Human-Readable Data
Context 2: Conclusion / A Case Study

Purpose: Apply insights to XML
Description: Critique of XML

-Based on a hierarchical structure which can cause problems if deep hierarchies are used
-No landmarks
-Spec provides no guidance on the usefulness of groupings
-Spec provides no guidance on how to choose tags
-Tag structure does not reference existing nesting structures like ( ) [ ]
-Tag structure wisely uses prepositions for keywords
-Only one level metadata (e.g. Not possible to embed comments in tags for holding meta-meta discussions about the purpose and meaning of tags and variables)
-A DTD describing the document structure correctly appears early in the document

Conclusion: The insights into human-readable data formats developed here allow us to perform non-trivial evaluations of existing practices