Research Designs Using Content Analysis

Klaus Weber
Northwestern University
klausweber@northwestern.edu

PDW "Content Analysis in Organizational Research"

Academy of Management
Montreal, 2010

A Definition

> "**Content analysis is a research technique for making replicable and valid inferences from texts [broadly conceived] to the contexts of their use.**"
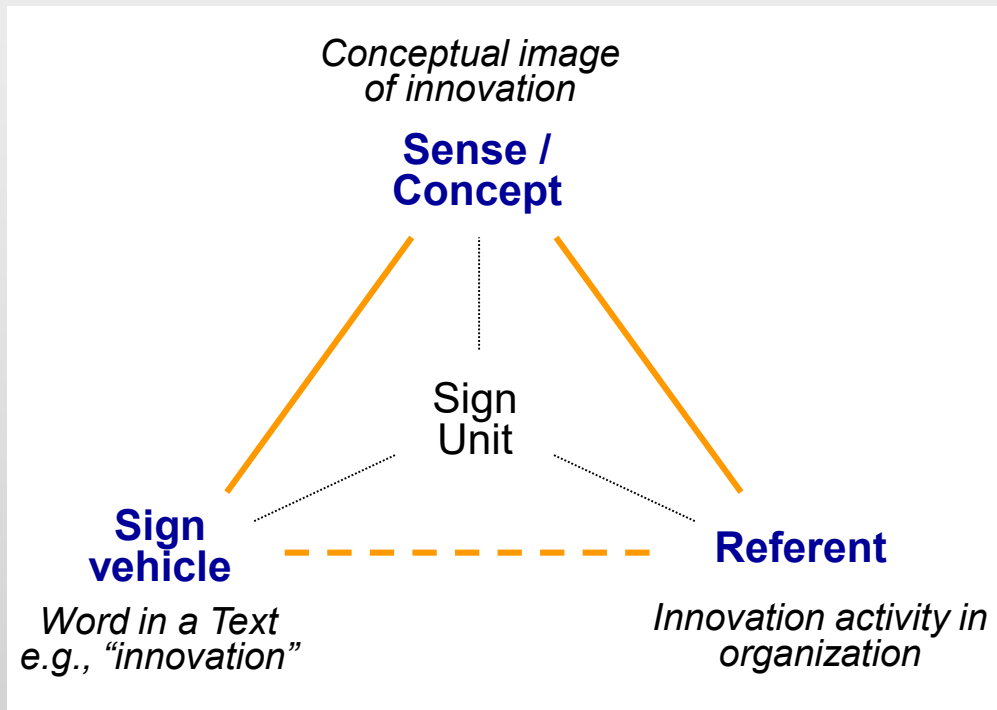>
> Krippendorff, 2004

- **Technique: Systematic (reliable, valid)**
  - Evidence based, transparency of procedure
- **Text: Data is generated in communication process**
  - Author, audience, intent, context matter
- **Inference: Interpretative in nature, inference about social reality**
  - Researcher, research question, theoretical constructs

- **Examples of how data is generated: Conversations, speeches, articles, open ended survey responses, interviews, images**

Outline

- **What's in a Text? A Semiotic Framework**

- **Generic Types of Textual Analysis**

- **The Process From Collecting Data to Presenting Results**

- **Sampling, Base Rates and Control Groups**

- **Validity and Reliability**

- **Software Support**

What's in a "Text"? A Semiotic Perspective

Conceptual image of innovation

Sense / Concept

Sign Unit

Sign vehicle

Word in a Text e.g., "innovation"

Referent

Innovation activity in organization

- <u>Text</u> = words/images arranged in order, but of interest are often *ideas* or *actions* that the words point to

- <u>The semiotic problem</u>: words, concepts, and referents do not correspond one-to-one (the good news: associations are usually conventionally defined in a particular social context)

-> Be clear about what the text analysis is getting at: Linguistic patterns, cognitive-cultural schemas, "facts"?

<u>Solutions to the semiotic problem</u>:

– Three generic sources of meaning and interpretation:

  – <u>Referential</u>: A word's meaning derives from its association with a referent or idea (e.g., categories, names)

  – <u>Relational</u>: Meaning derives from a signs position to other signs (e.g., association, opposition, grammar, plots)

  – <u>Contextual</u>: Meaning derives from the communication context (who created the text, for whom, when, why)

What Content Is Analyzed

Inductive Theory Building

- Holistic interpretation (-> themes, mechanisms, taxonomies)

Quantitative Content Analysis

- Small text units in isolation, e.g. categories (-> frequencies, trends, etc.)

Semantic Analysis

- Relationship between content units, e.g. associations and grammar
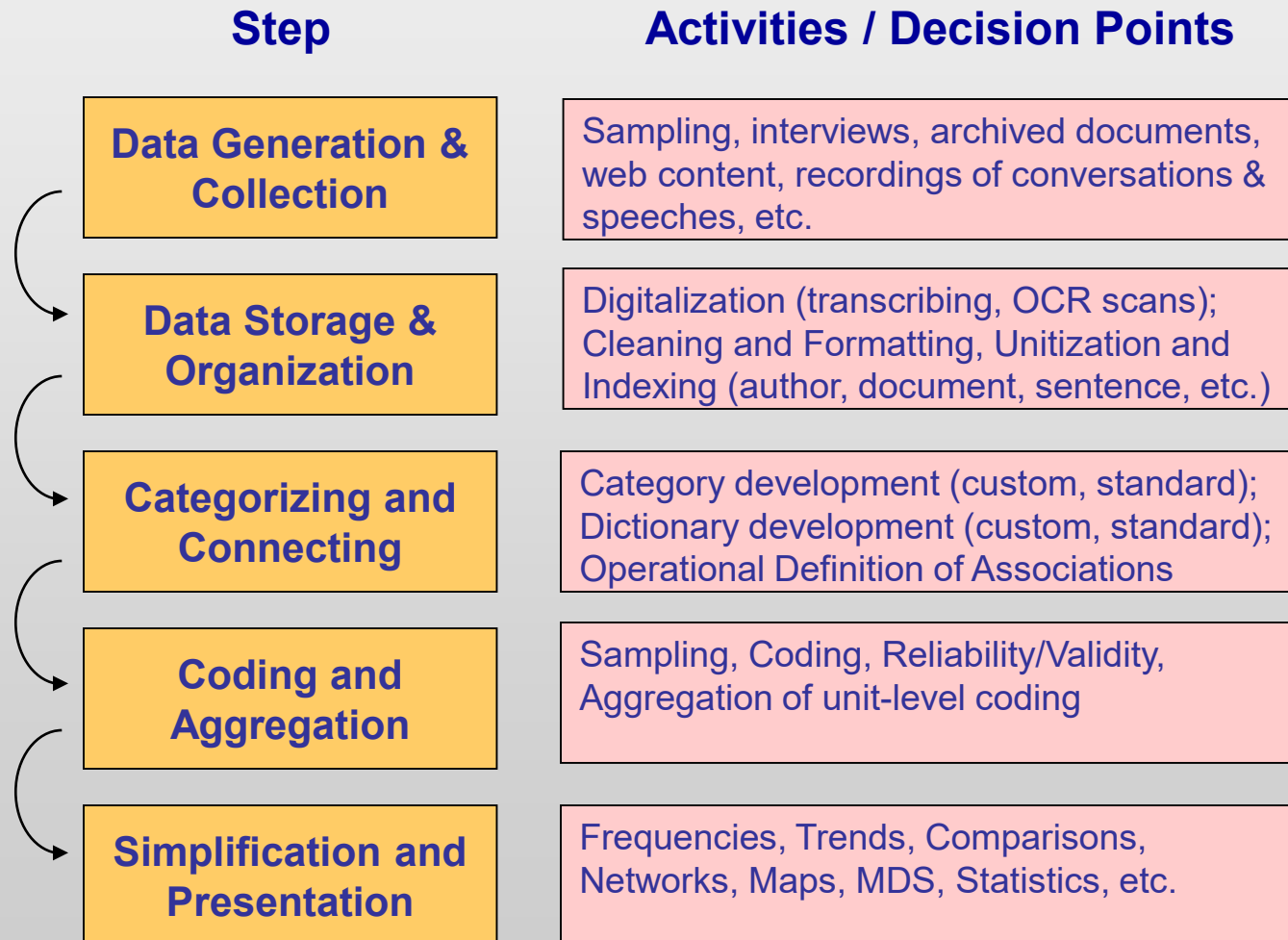 (-> scripts, networks of associated concepts, causal maps, narratives)

Computational Linguistics

- Algorithm based identification of structure, e.g. latent semantic clusters
  (-> analogue: factor and cluster analysis; problem: interpretation)

Discourse Analysis

- Several texts, e.g. regimes of interpretation
  (-> broad ideologies, institutional myths and political contradictions)

The General Process

Step	Activities / Decision Points
Data Generation & Collection	Sampling, interviews, archived documents, web content, recordings of conversations & speeches, etc.
Data Storage & Organization	Digitalization (transcribing, OCR scans); Cleaning and Formatting, Unitization and Indexing (author, document, sentence, etc.)
Categorizing and Connecting	Category development (custom, standard); Dictionary development (custom, standard); Operational Definition of Associations
Coding and Aggregation	Sampling, Coding, Reliability/Validity, Aggregation of unit-level coding
Simplification and Presentation	Frequencies, Trends, Comparisons, Networks, Maps, MDS, Statistics, etc.

Sampling, Base Rates and Control Groups

- **Sampling "authors"**
  - **Most common**

- **Sampling texts**
  - **No need to code every document (in most instances)**

- **Sampling coding units**
  - **No need to code every paragraph / sentence (in most instances)**

- **Base rates of word or category occurrence**
  - **Random is fine but often not most realistic**

- **Control groups**
  - **Corpus linguistics**
  - **Closely matched other texts**

Validity and Reliability of Inference

- **Validity: external (generalization), internal (constructs and causality)**

- **Reliability: replication over time, across individuals**
  - **Multiple researchers for category development**
  - **Applicability of generic categories and dictionaries**
  - **Documentation (code book, dictionary)**
  - **Reliability statistics: Don't use percent agreement, avoid Cohen's kappa; use statistics that account for chance agreement between coders: Krippendorff's alpha, Scott's pi**

Software Support Options: Some Packages

Type of Software	Examples of popular software	Functionality Storage, retrieval	Functionality Developing and linking categories	Functionality Automated content coding	Functionality Mapping, display of coded data	Functionality Quanti-fication, statistics
Theory Building Support	ATLAS.ti, Ethonograph, Kwalitan, MaxQDA, NVivo	Yes (best for smaller volumes)	Yes (main focus)	Some (best for smaller volumes)	Some (mostly basic)	Little (export to other software)
Coding Support	Diction, TextQuest, LIWC Yoshicoder WordStat	Yes	Little	Yes (main focus, efficient for high volume)	Little	Little (export of other software)
Mapping	AutoMap, Decision-Explorer, UCINet	Some	Little	Yes	Yes (main focus)	Some (e.g. concept centralities)
Text Mining	TextAnalyst, SAS plug-in, SPSS plug-in WordStat, TAKMI	Yes (especially for large volumes)	Little	Yes	Some	Some (e.g. built in algorithms)

A Few Key Questions To Ask

- **What does the "content" represent?**

- **How were the data generated?**

- **Are categories and measurements valid and replicable?**

- **What is the base rate of "content" occurrence?**

- **What inferences / interpretations can I actually make?**