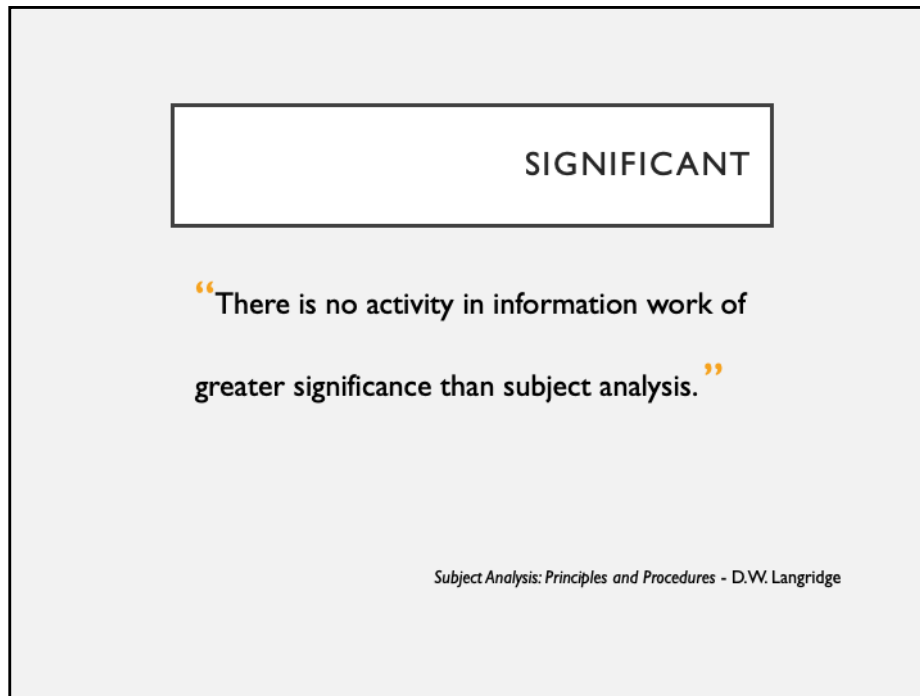


Subject Analysis

LIS 530
Winter 2019



This is the first sentence of one of my favorite information science books, Subject Analysis: Principles and Procedures. I also thought it would be a good way to get your attention early in the morning. So, let's talk about why subject analysis might be significant.....

[Classification Research Group](#)

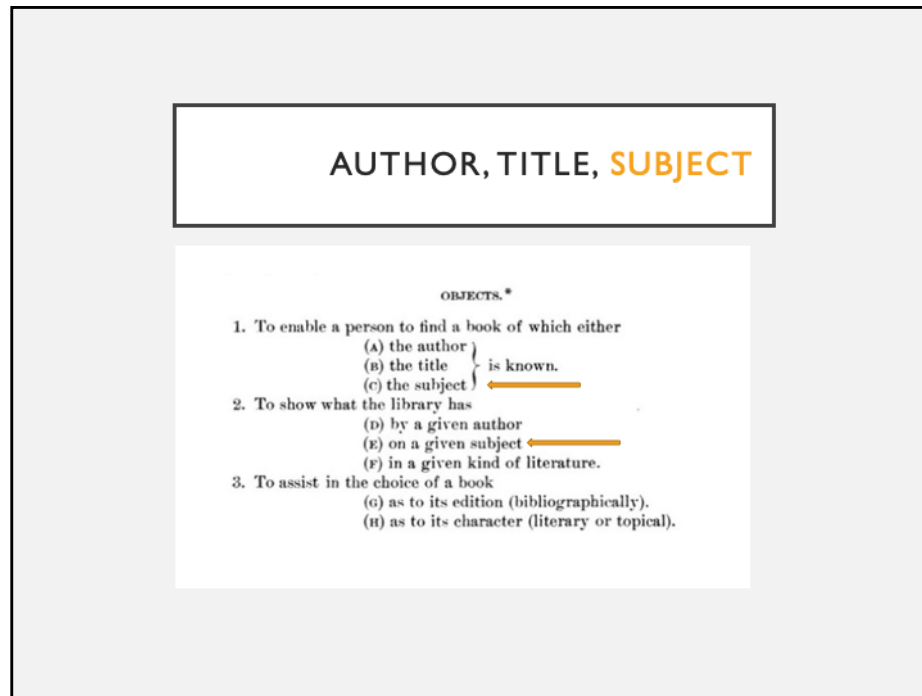
LIKE INFORMATION TOGETHER

“The essential and defining objective of a system for organizing information, then, is to bring essentially like information together and to differentiate what is not exactly alike.”

The Intellectual Foundation of Information Organization – Elaine Svenonius, p. 11

This quote from Svenonius follows from her discussion of Panizzi.

Question for class: Why do we want like information together?

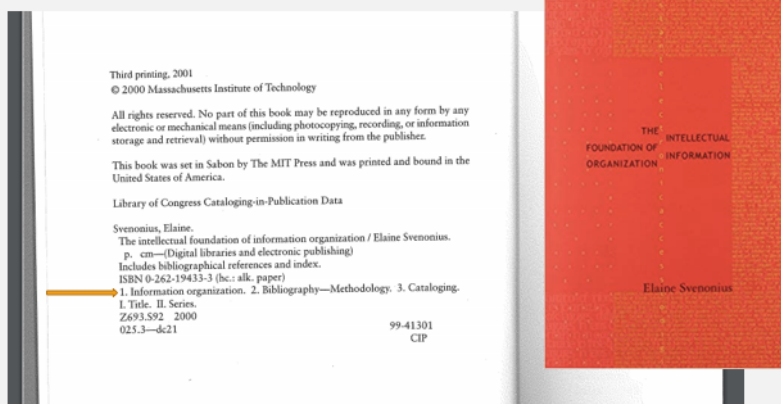


These are Cutter's rules from the 1876. They're from the first column in Wan-Chen's table

Three main attributes that we use to organize and find information.

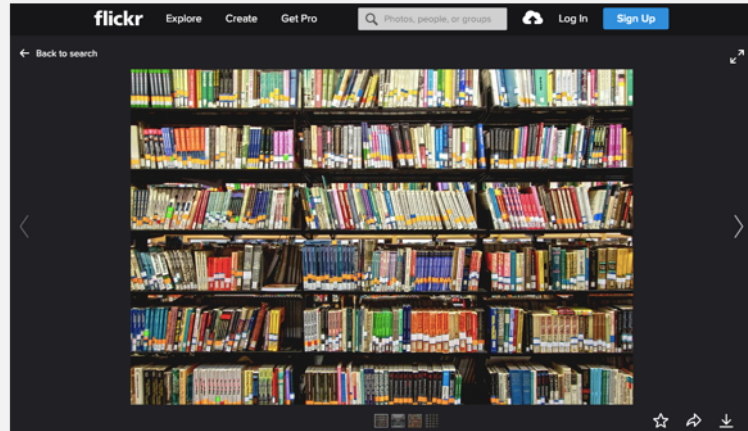
- Author, aside from authors with multiple names and corporate authors, are relatively straightforward and not subjective.
- Similar for Title, which generally goes with the title provided by the author and publisher on the title page. (This is not nearly as simple for non-bibliographic materials.) Again, this is objective; it doesn't require the classifier to make subjective choices.
- But then we have subject. What does "subject" mean to you? What the item is about? What topics it covers?

SUBJECT: AN EXAMPLE

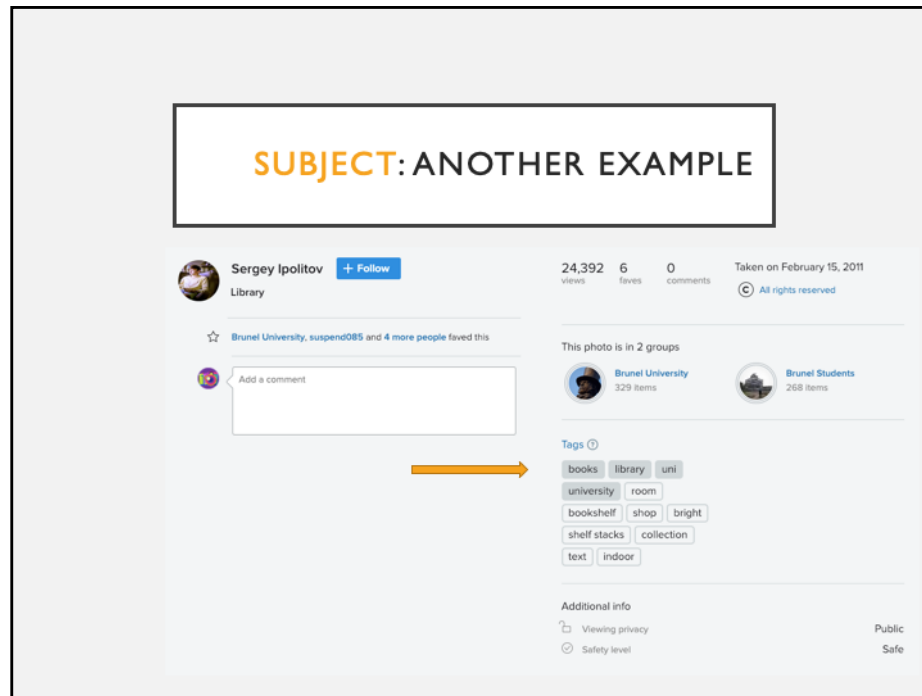


Here's the result of subject analysis work that you're probably familiar with in the frontmatter of a book.

SUBJECT: ANOTHER EXAMPLE



What is the subject of this photograph? What do you see? What is it about?



Here's the metadata associated with this photo. We see the photographer, the title, and a bunch of other stuff. And we see these tags. Are these the "subject" of the photograph? Would you use "room" or "indoor"?

Note that you can click on the "books" tag and see other photographs that are tagged "books." This is a way of supporting colocation. Like items – in this case photographs about the same thing – can be found together. This aids retrieval. We do a similar thing with books and other materials, where we can access them by subject.

EXERCISE



What is the content of the article – what does it *contain*?

What is the subject of the article – what is it *about*?

So, now that we've seen a couple examples but not talked in depth about subject analysis, let's do some subject analysis!

Instructions: Open the article *How Language Influences Emotion* and ask yourself these two questions. Keep your answers separate. We'll talk about how your answers and your approaches differ.

Follow up questions:

What is the content of the article? Hit on topics, names, time, form.

What was your approach to determining the content? Did you read the whole article from start to finish? Did you pay attention to specific parts of the article? Metadata?

What is the subject of the article?

How do you think subject is different from content?

What was your approach to determining the subject?

One distinction to make – You can do content analysis by scanning the document for key words and metadata fields. To do subject analysis, you need to do

CONTENT ANALYSIS

More general than subject analysis.

Includes:

- Topics (What's discussed and communicated?)
- Names (Who's mentioned?)
- Time (Is a time period or point in time mentioned?)
- Form (Is content communicated because of the form it is in – inventory chart, annual report, white paper, etc.?)

So, the first question from our exercise what CONTENT analysis, which is more general than subject analysis. It includes more things – there are few restrictions on what you can consider content analysis.

This list of what is included comes from Taylor.

SUBJECT ANALYSIS

More specific than content analysis.*

Here we want to know specifically what the document is *about*. What is the subject of the document?

Subject analysis has been debated frequently in research literature. Let's look at some approaches...

* Note that Langridge, who I cited at the beginning of class and who we'll cover again soon sees these terms are largely interchangeable, prefers content analysis, and uses subject analysis because of historical precedent.

SUBJECT ANALYSIS: WILSON

1. Identify *Cast of Characters*.
2. Perform one or all of these methods:
 - Purposive Way
 - Figure-Ground Way
 - Counting References Way
 - Appeal to Unity Way

Two Kinds of Power: An Essay on Bibliographic Control – Patrick Wilson

Introduce Wilson and Cast of Characters.

WILSON: CAST OF CHARACTERS

“Let us ask a person to write down, for the sentences given, a list of **things named**, **mentioned**, or **referred to**, and also of **concepts employed** in addition to those employed in the naming, mentioning, and referring to things. Let us call such lists, Casts of Characters of sentences. ”

Two Kinds of Power: An Essay on Bibliographic Control – Patrick Wilson

Introduce Wilson and Cast of Characters.

WHERE TO LOOK



Title
Table of Contents
Section Headings
Preface/Introduction
Foreword
Abstract/Summary
Index

Okay, this isn't specific to Wilson, but when we do something like look for the cast of characters, do we read a whole book from cover to cover? Probably not. We can look in some places where we can expect high value information.

Anywhere else that you expect to find lots of useful information?

SUBJECT ANALYSIS: WILSON

1. Identify *Cast of Characters*.
2. Perform one or all of these methods:
 - Purposive Way
 - Figure-Ground Way
 - Counting References Way
 - Appeal to Unity Way

Two Kinds of Power: An Essay on Bibliographic Control – Patrick Wilson

Okay, so once we've collected all of the information about the document, we can work on understanding what it is about – its subject. Wilson provides 4 ways to do this and we'll look at—and critique--each of them. When you do subject analysis work, you can choose one or more of these methods. If one doesn't work for your document, don't use it.

WILSON: **PURPOSIVE**

What was the creator's aim or purpose with this document?



- Might be difficult to interpret the creator's intent unless they make an explicit statement.
- Statement might not be in the text and creator's often make multiple contradictory statements.
- Some works become about something other than what the creator intended.

Two Kinds of Power: An Essay on Bibliographic Control – Patrick Wilson

This seems great, right? The author of a book tells you what it is about, and the author is the best expert, right? But it isn't necessarily that easy.

- The creator doesn't always state the purpose or that statement is hidden in an interview that you don't have access to.
- The creator might state multiple contradictory purposes for a work or document.
- Popular culture, historical perspective, or specific communities might make it so that a work is generally viewed to be about something other than what the creator intended.

Important contrast: When we list author and title, we generally accept the purposive way without trouble. For subject, it isn't so easy.

WILSON: FIGURE-GROUND

What stands out? What is the central figure?

- The central figure could be a person, a concept, etc. It could be a group if one single entity doesn't stand out.
- "Stands out" is open to interpretation
- What's "important" can depend on the user. For example, public library vs. medical library.
- What's "important" can change over time or even from day to day.



Two Kinds of Power: An Essay on Bibliographic Control – Patrick Wilson

Look, more penguins! I heard that you like penguins. 😊

Saying that a document is about what stands out as central can be useful. What is the subject of a Dewey biography? What is the subject of my daughter's book about storms, temperature, precipitation, clouds, barometric pressure?

We want to be unbiased when we do subject analysis, but identifying what stands out is likely to introduce some bias.

History can change our view of what's important.

And so can our day-in and day-out experiences. What you think is important one day might seem less important to you a few days or a week later.

WILSON: COUNTING REFERENCES

What is mentioned the most times?

- Also called the “Objective Way” because it is just a numerical count.
- But the real subject of a document might never be explicitly stated.
- Synonyms can also create problems with counting.



Two Kinds of Power: An Essay on Bibliographic Control – Patrick Wilson

If something is mentioned a lot, it must be important, right? Whatever is mentioned most is the subject of the document, then.

This approach can be useful, at least as a first cut.

But often times key topics are not mentioned explicitly because they are implied.

Or there's a lot of background build-up to get to the actual subject. But the subject itself isn't listed as many times.

Or synonyms can throw off your count.

For large works, counting without a computer could be quite difficult.

WILSON: **APPEAL TO UNITY**

What is the central theme? What unifies the document?



- Also called the “Cohesion Way” because it is about what makes the document cohesive.
- Might require significant subject expertise to understand what the unifying principle is for some documents.
- Might require you to know what it omitted.

Two Kinds of Power: An Essay on Bibliographic Control – Patrick Wilson

What is the theme that runs throughout all of the topics in a document?

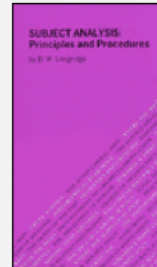
Good example: Wan-Chen, Joe, I, and a few other students are working on a book with the working title “Disciplines of Description.” We’re expecting contributed chapters about bibliographic description, museum description, species description, maybe dance notation – lots of different topics and field, but the unifying principle is description.

Omitted example: Suppose that I have a cookbook with only main dishes--no sides, no desserts, no drinks, etc. If you aren’t used to looking at cookbooks, you might not realize that a unifying principle of that book is that it is all about main dishes unless maybe the title was “Main Dishes!”

SUBJECT ANALYSIS: LANGRIDGE

Examine the document and ask:

1. What is it about?
2. What is it?
3. What is it for?



Subject Analysis: Principles and Procedures - D.W. Langridge

Langridge wants us to ask three questions about a document, not confuse them with each other, and use the synthesis of our answers to understand the subject.

LANGRIDGE: WHAT'S IT ABOUT?

This question is about the topical aspects of the work.

<i>Introduction to Science</i>	<i>History of Science</i>
Covers: Matter, energy, life, etc.	Covers: Names of scientists and their activities.
What is it about? Science.	What is it about? Science.

Subject Analysis: Principles and Procedures - D.W. Langridge

We also call forms of knowledge disciplines. DDC, for example, is largely structured around disciplines. The Burke reading for today's abstract is about how we've built categorization and classification schemes around disciplines.

LANGRIDGE: WHAT IS IT?

“The way in which the phenomena...are treated....forms of knowledge, like science, history, and philosophy.”

<i>Introduction to Science</i>	<i>History of Science</i>
Matter, energy, life, etc.	Names of scientists and their activities.
What is it about? Science.	What is it about? Science.
What is it? Science.	What is it? History.

Subject Analysis: Principles and Procedures - D.W. Langridge

We also call forms of knowledge disciplines. DDC, for example, is largely structured around disciplines. The Burke reading for today’s abstract is about how we’ve built categorization and classification schemes around disciplines.

LANGRIDGE: WHAT IS IT FOR?

This question is about audience, which is often implied but rarely stated explicitly.

“A book about X is for those interested in X.”

<i>Introduction to Science</i>	<i>History of Science</i>
Matter, energy, life, etc.	Names of scientists and their activities.
What is it about? Science.	What is it about? Science.
What is it? Science.	What is it? History.
<i>Intro to Science for Politicians</i>	<i>History of Science for Physicists</i>

Subject Analysis: Principles and Procedures - D.W. Langridge

Langridge is somewhat unique in considering this question a core part of subject analysis.

That might be because identifying the audience is not always easy or, more often, because the audience is simply those interested in the book, as reflected in Langridge’s tautology above.

Examples from Langridge:

- “Economics for industrial managers”
- “Statistics for engineers”
- “Psychology for nurses”

But understanding audience can be critical for putting like materials together and aiding retrieval.

SUBJECT ANALYSIS: LANGRIDGE

Subject analysis should be **separate** from classification.

You should use **natural language** and reconcile your analysis with a controlled vocabulary or index like DDC or LCSH in the next step.

But it rarely works that way in practice. These steps are frequently combined.

Subject Analysis: Principles and Procedures - D.W. Langridge

Why do you think these steps are frequently combined?

Why do you think that Langridge advocates for separating them?

Reinforce where we are in the process. Subject analysis is an initial evaluation. We're not trying to fit with a scheme yet. We're trying to understand what the document is about. There's a whole bunch of work left to classify, index, catalog, etc.

SUBJECT ANALYSIS: FACETS

What if a document's subject is based on a combination of things?

Langridge was hinting at this by asking three questions.:

<i>Introduction to Science</i>	<i>History of Science</i>
Topic: Science	Topic: Science
Form: Science	Form: History
Audience: Politicians	Audience: Physicists

We can treat those questions as facets: Topic, form, audience

If we combine these facets, we're building a faceted classification scheme. We'll talk much more about faceted classification later in the class and compare it to enumerative classification.

Let's just take a quick look at it here today.

A couple notes:

- Langridge was a member of the Classification Research Group in the UK. Ranganathan, who developed faceted classification, visited and worked with the Classification Research Group.
- The differences between faceted and enumerative classification (and the differences between different schemes) reinforce Langridge's assertion that subject analysis should be separate from classification.

EXAMPLE OF FACETS

With a classification of cars, e.g.:

Color	Brand	Transmission	Type
Red	BMW	Manual	Sedan
Blue	Volvo	Automatic	Wagon
Green	Saab		SUV
Yellow	Volkswagen		Truck
Black			

Here's an example of facets for cars.

FACET ANALYSIS

- Facet Analysis: Breaking down the subject of a document into component parts.
- Facets: The component parts of a subject.
- Facets represent Fundamental Categories.

FACETS: KAISER

Julius Otto Kaiser was the first to propose the use of facets in 1911.

Kaiser had three fundamental categories or facets:

- Concretes – nouns
- Processes – verbs and adjectives
- Place – mostly countries

Kaiser (1868-1927) was born in Germany but traveled widely. He worked as a special librarian and indexer in the US and the UK.

- Concretes/Nouns could be general things-in-the-world or domain specific terms. Kaiser distinguished between what he called moveable commodities (hardware, paper, money) and immoveable commodities (land, rivers, etc.).
- Processes are both conditions of concretes (adjectives) or actions associated with them (verbs)
- For place I say “mostly” countries because Kaiser also used geographical regions that were not politically unified.

Kaiser went back and forth on whether his system used triads or dyads. Place or country wasn't there initially and was sometimes considered optional. Concretes were considered the main facet, but Kaiser even played with dyads that did not include concretes.

Note the similarity to the geography table in the DDC. What do we think of the importance placed on place here now that travel and communication are much easier than in 1911?

What values would we assign to these facets for Elaine Svenonius's book? [Concretes: Information] – [Processes: Organization] – [Place: United States?]

FACETS: RANGANATHAN

S. R. Ranganathan (1950s-1972) had five fundamental categories or facets in his analytico-synthetic Colon Classification:

- Personality
- Matter (i.e., Material or Property)
- Energy
- Space
- Time



Ranganathan was from India. His colon classification is widely cited as **the** major development in faceted classification. He's also Joe's favorite LIS scholar.

It says analytico-synthetic because it is based on analyzing (subject analysis) and then synthesis (combining the facet values that you identify during analysis).

Why isn't colon classification more popular and widely used?

- Ranganathan was from India and, while he traveled to the UK and worked extensively with the Classification Research Group, he was an outsider to much of the LIS scholarship.
- May considered his system too complicated.

Colon classification is more widely used in India.

So, let's unpack PMEST, because I'll admit that I find it a bit cryptic and I bet you do do.

FACETS: RANGANATHAN

- **Time:** Are the resources about a certain time or period of time?
- **Space:** Are the resources about a certain place?
- **Energy:** Are there a concepts of a certain action on the main subject of the resource?
- **Matter:** Are there a concepts of a certain material or property of the main subject of the resource?)
- **Personality:** What' s left over? What is the substance of the resources? What is their main subject (minus MEST)?

Note that I've flipped the order here. P has to come last. It is both a catch-all and the main topic.

Okay, let's look at each of these....

RANGANATHAN: EXAMPLE

Seed-borne biocontrol agents for the management of rice sheath rot caused by *Sarocladium oryzae* (Sawada) W. Gams & D. Hawksw.

Author(s) : Gopalakrishnan, C.; Yallanarajan, N.

Author Affiliation : Department of Plant Pathology, Centre for Plant Protection Studies, Tamil Nadu Agricultural University, Coimbatore 641 003, Tamil Nadu, India.

Author Email : gopal@yallanarajan.com

Journal article : *Journal of Biological Control* 2006 Vol.20 No.2 pp.197-204 ref.14

Abstract : All the antagonist isolates (10 isolates of *Pseudomonas fluorescens*, 15 of *Bacillus subtilis*, and 3 of *Trichoderma viride*) obtained from rice seeds showed in vitro inhibitory effect on *S. oryzae*, ranging from 10.90 to 82.18%. *B. subtilis* isolate 9 was found to be highly effective in inhibiting the mycelial growth of *S. oryzae* by 82.18%. The biological control agents significantly reduced *S. oryzae* seed infection and increased seed germination and seedling vigour. Field trials conducted at four different locations (western, southern, northeastern and Cauvery delta) in Tamil Nadu, India, during rabi 2004 indicated that the seed-borne *S. oryzae* in rice could be successfully managed by seed treatment (at 10 g/kg), followed by foliar application (at 1%) at boot leaf stage with *B. subtilis* or *P. fluorescens*.

ISSN : 0021-9304

Record Number : 20073041316

Publisher : Society for Biocontrol Advancement

Location of publication : Bangalore

Country of publication : India

Language of text : English

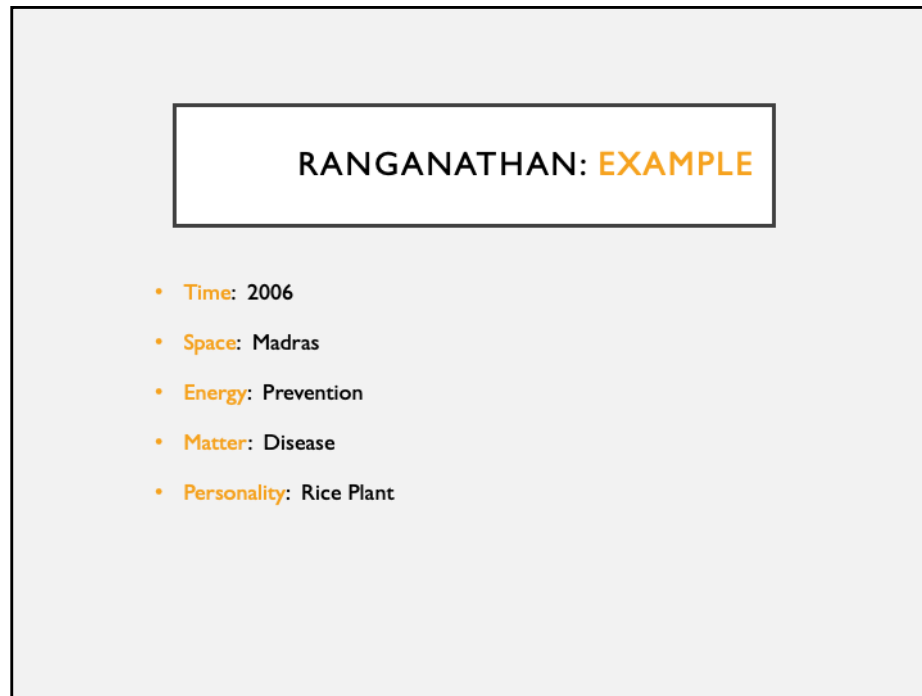
Language of summary : English

Indexing terms for this abstract :

Organism description(s) : *Bacillus subtilis*, fungi, *Oryza*, *Oryza sativa*, *Pseudomonas fluorescens*, *Sarocladium oryzae*, *Trichoderma viride*

Descriptor(s) : biological control, biological control agents, foliar application, fungal antagonists, fungal diseases, growth, mycelium, natural enemies, pathogens, plant disease control, plant

Here's an example. This is an interesting example because I was working with Joe's slides from previous years. I couldn't find the document that he used in his example, but I had the values that he assigned for each facet. I worked backward and found this article, which might just be the one that he used.

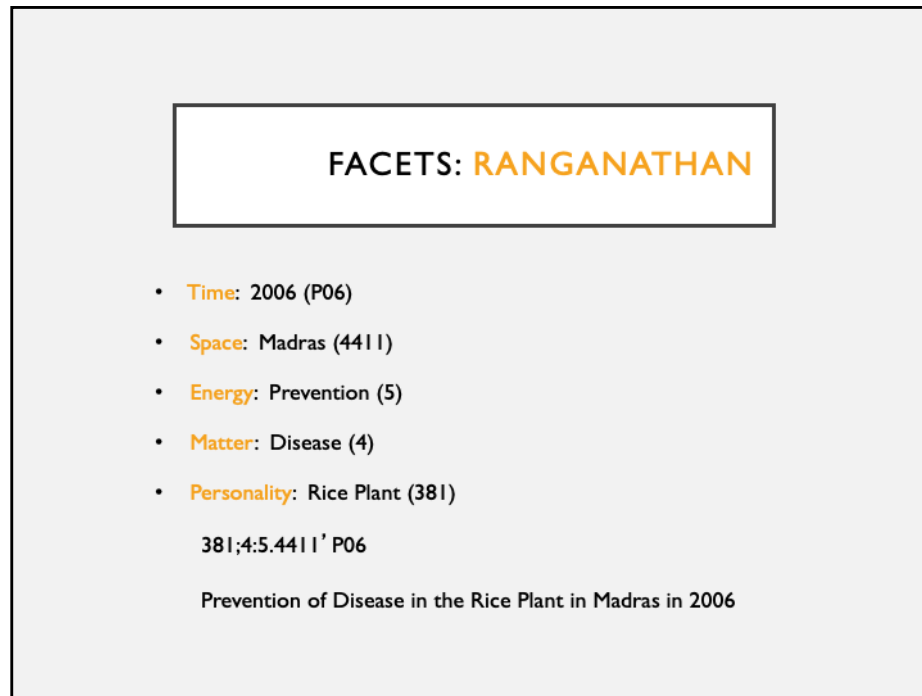


Here's an example:

- Time is straightforward: 2006 is the publication date.
- Space is a bit trickier. I assume that Joe had Madras listed here because Madras is in Ranganathan's vocabulary, but Madras is no longer a state in India. It is Tamil Nadu, as listed in the article metadata.
- Energy is what is being done. We don't see the word "prevention" anywhere in the metadata or the abstract, but it does a good job of summarizing what they are doing – they are preventing rice sheath rot. The word in the title, "Management" is more ambiguous.
- Matter is tricky. Should it be rice? Or disease? It is disease for a couple reasons: 1) The prevention is an action on the disease and 2) When someone needs to retrieve this information, they're more likely to be interested in Rice, Disease Prevention than Disease Prevention, Rice. That's a judgment call, though.
- - And now we're down to Rice (specifically the plant, not the food) because that's the main topic and because that's what's left.

Worth noting: Some of these terms do not show up in the metadata and abstract. That's partly because they are in Ranganathan's controlled vocabulary. But it is also

because you need to do some analysis when you do subject analysis. It isn't just finding keywords.



We're not going to get into classification schemes and notation, but you can get a quick sense here of how these facets would be combined in a scheme.

Note that the organization structure is based on a combination (a synthesis) of the five facets. The first number we see there is for P, which is the primary way to find the document.

SUBJECT ANALYSIS: SUMMARY

- Subject analysis is important for putting like information together and supporting a key mode of information retrieval.
- Subject analysis is more specific than content analysis.
- **Wilson:** Purposive, Figure-Ground, Counting References, Unity
- **Langridge:** What is it about? What is it? What is it for?
- **Facets:** Breaking down subject into component parts.
- **Kaiser:** Concreteness, Process, Place
- **Ranganathan:** PMEST

Langridge: Also that subject analysis should be separate from classification.