Computer Science and Application 计算机科学与应用, 2020, 10(11), 2058-2063
Published Online November 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.1011217

Hans汉斯

基于时空场景的话语元数据研究

关　琳[1,2]

[1]江苏警官学院公安管理系，江苏　南京
[2]南京大学中国智库研究与评价中心，江苏　南京
Email: 17590324@qq.com

收稿日期：2020年11月4日；录用日期：2020年11月19日；发布日期：2020年11月26日

摘　要

话语承载思想，为推动话语资源的科学组织，促进话语的传播、宣传、研究和阐释，文章对话语元数据开展研究，梳理了话语资源现状，确立话语元数据设计的核心性、扩展性、抽象性和丰富性原则，并基于这些原则建立了一套基于时空场景的话语元数据标准。

关键词

话语，元数据，时空场景，数据库

Research on Discourse Metadata Based on Space-Time Scene

Lin Guan[1,2]

[1]Public Security Administration Department, Jiangsu Police Institute, Nanjing Jiangsu
[2]China Think Tank Research and Evaluation Center, Nanjing University, Nanjing Jiangsu
Email: 17590324@qq.com

Received: Nov. 4th, 2020; accepted: Nov. 19th, 2020; published: Nov. 26th, 2020

Abstract

Discourse contains thoughts to facilitate the scientific organization of the discourse resources, promote the spread, propaganda, research and interpretation of discourse, the article metadata research, summarize the discourse resources present situation, establish discourse metadata design core, extensibility, the principle of abstraction and richness, and based on these principles to establish a set of discourse metadata standard based on time-space scenario.

文章引用: 关琳. 基于时空场景的话语元数据研究[J]. 计算机科学与应用, 2020, 10(11): 2058-2063.
DOI: 10.12677/csa.2020.1011217

Keywords

Discourse, Metadata, Space-Time Scene, Database

1. 引言

随着移动互联网的普及和信息技术的快速发展，网站、APP 等数字信息资源已成为人们主要的知识来源渠道。在党的思想理论建设领域，数字资源的科学建设和高效利用关乎宣传思想政治工作的成效。党的十九次全国代表大会指出习近平新时代中国特色社会主义思想(下文简称“习近平思想”)是全党全国人民实现中华民族伟大复兴的奋斗指南。当前，研究“习近平思想”是解读我国当前各类政策、顶层设计的基础和重要切入点。承载“习近平思想”的话语是这项研究的原始对象和起点。

本文通过建立“习近平思想”话语元数据(以下简称“话语元数据”)核心标准，为构建思想理论领域话语知识库，整合和开发利用话语资源，突破当下以全文本为主的数据库资源模式奠定基础，并助力于宣传思想政治工作，推动理论思想的研究和传播。

2. 话语元数据的概念和功能

“话语”是对象在各类正式、非正式场合，以官方、半官方的渠道，产生的承载思想的书面或口语化的文献记录。元数据是描述信息或数据资源自身属性和特征的数据，在信息资源组织和管理过程中发挥重要作用[1]。话语元数据是一套通用的，独立于平台的，建立在目录学理论基础上，能够描述话语文本数据属性的模型。建立话语元数据旨在识别话语资源并追踪话语资源在时空场景语境下的发展和变化，从而支撑话语研究、解读和宣传。

话语元数据是通过独立于平台、规范统一的方式对话语文本数据的模式予以描述，通过一套资源模型结构来表达话语的通用信息[2] [3]。话语元数据的功能体现在以下几个方面：

第一，话语元数据独立于当下各类“习近平思想”资源数据库平台工具，提供的是一套基于时空场景的话语数据库模型基础建构方案，可以通过编码将话语元数据转换成话语数据库；

第二，话语元数据为现有系统提供对照参考模型，可为现有“习近平思想”宣传数据库系统提供对照和完善参考。话语元数据并不包含数据库或平台特性，提供一种针对“习近平思想”话语资源的通用的元数据描述，助力于该领域资源的科学组织和合理利用。

第三，话语元数据将话语产生的时间、空间和场景抽象为元数据模型，用于还原在线话语产生和发展的时空场景，为开展“习近平思想”话语研究提供拓展空间。

3. 话语资源构成现状

围绕“习近平思想”话语原文，衍生出如图 1 所示六大类不同类型的文献，这些文献按照用途被整合在不同的资源平台上。以宣传平台为例，目前收录“习近平思想”文献的权威数据库有三个。分别是由中央网信办指导、人民网·中国共产党新闻网建设的“学习路上——习近平总书记系列重要讲话大型网络数据库”(2014 年建成) [4]；由人民出版社开发建设的“中国共产党思想理论资源数据库”(2010 年建成)；以及由中宣部(2018 年建成)推出的“学习强国”学习平台[5]。除这三家外，还有大量的类似的习

近平话语数据库工具或网页，广泛存在于各类党建数字资源中。这些资源的创建为学习和研究习近平话语带来了便利。

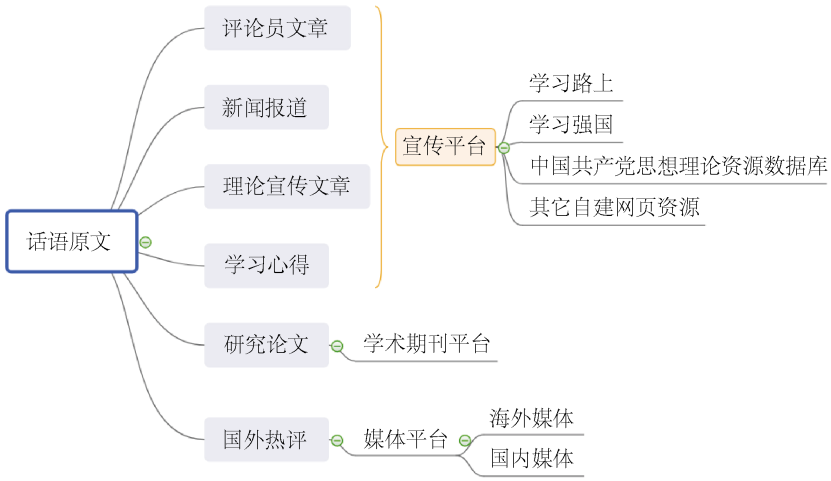

Figure 1. The structure of discourse resources and the status quo
图 1. 话语资源构成与组织现状

但是，从为数众多的分散的数据库中筛选查询自己需要的资源，并获取到具体时空场景下的话语文本，却并不容易。主要原因在于这些数据库，在功能上缺乏基本的文本统计、计量、分析手段；在内容上以衍生文献为主，多为新闻通稿而非话语原文；从信息的组织形式上看仅仅完成了各类文件的分类、归纳和保存，粗粒度数据无法支持话语文本研究，如基于时空场景的词频分析、文本挖掘等场景的需要。因此，只有通过建立时空场景话语元数据，将来自不同平台的话语资源充分整合，才能够建立起“习近平思想”话语的全面图景。

4. 话语元数据设计遵循的原则

为更好的适应“习近平思想”宣传和研究的现实需要，确保话语元数据的科学性和通用型，在话语元数据设计过程中应遵循四项原则：

(一) 核心性原则

话语元数据的设计应充分遵循核心性原则。话语元数据应作为思想理论宣传资源的核心元数据，应独立于具体系统的设计，以保证元数据的独立性。同时，应能够以话语元数据为核心，在其基础上向不同类型的话语资源扩展差异化的元素或元素子集。

(二) 扩展性原则

话语元数据的设计应充分遵循扩展性原则。话语元数据设计应具备前瞻性，为将来话语资源的丰富和发展预留空间，元数据应具备向后兼容能力。对于难以获取的字段，在设计过程中可适当留白(暂不取值)，以便留待将来使用。

(三) 抽象性原则

话语元数据的设计应充分遵循抽象性原则。抽象性体现在包括需求分析、研究设计和模型建立等，话语元数据设计的各个环节。在需求分析阶段必须对话语资源使用场景充分抽象，才能兼顾话语资源当下的组织利用和未来的深度挖掘；在研究设计阶段应充分调研现有话语资源平台的数据模型，充分抽象使话语元数据充分适应现有各类系统；在模型建立阶段应对话语资源发展趋势充分抽象，使话语元数据

模型能够适应未来系统建设和资源利用。

(四) 丰富性原则

话语元数据的设计应充分遵循丰富性原则。话语载体的形式随时代发展而日趋多样[6]。话语研究的场景也随着技术的进步而愈发复杂和深入。这些客观现实都不断为话语资源的组织提出新的要求。只有遵循丰富性原则，使话语元数据尽可能多的囊括资源的描述维度，才能延续话语元数据的生命周期。

5. 基于时空场景的话语元数据核心标准设计

基于时空场景的话语元数据核心标准是这项研究的最终落脚点。通过对话语元数据的需求开展分析，结合前期数据库研究基础，开展话语元数据标准设计。

(一) 基于时空场景的话语元数据需求分析

基于时空场景的话语元数据需求分析，从话语资源的保存视角和利用视角两个维度展开分析。

1. 推进话语资源的妥善保存

为了科学的组织话语资源，呈现话语产生和发展的真实状态，需要对话语资源自身属性信息和时空场景等各种相关信息妥善保存，以保证话语及其时空场景要素的完备。因此，话语资源的元数据应既包括话语本身的描述，也包括话语产生时空场景主题的描述。

2. 推进话语资源的高效利用

对话语进行描述、分类、保存和管理的最终目标在于利用话语资源开展研究和宣传工作。就话语研究而言，用户的检索维度千差万别，为提高检索效率，要求元数据能够从不同检索维度进行资源描述。就话语宣传而言，为更好的阐释话语，需要重构话语产生和主题演化的时空场景，这就需要元数据能够描述资源的静态特征和动态变化。

(二) 自底向上的元数据设计思路

基于前期构建的“基于时空场景的话语文本数据库”，将已经收集到的话语资源及其时空场景数据，再次提取，并优化元数据和原模型，提炼出基于时空场景的话语元数据核心标准。

(三) 基于时空场景的话语元数据核心标准

基于前述话语资源现状的梳理，本文遵循前述话语元数据设计原则，立足于话语元数据的现实需求，建立了话语资源元数据核心标准，如图 2 所示。在该核心标准中，元数据既包含话语文本的描述类数据，也包含话语产生的时空场景描述类数据。

其中话语文本描述类数据包括语言、主题分类、摘要、来源、责任者、题名、类型等七类元素；时空场景描述类数据包括时间、地点、受众、场景等四类元素。这 11 类元素共同描述话语文本的核心属性、话语产生和演化的时间、空间和场景等属性。

再将 11 类元素分成若干子集，以场景元素为例。话语的表述方式与场景密切相关，可将场景元素的子集做进一步扩充，包括出访、会见、致函、致电、考察、活动、会议、讲话等。其中，致函、致电为书面语形式，出访、会见多为外事场景，考察和会议类场景的话语常常以新闻通稿为主要来源，活动和讲话场景的话语则以活动讲话全文为主。

在核心性方面，话语元数据标准可以嵌入到现有三大权威话语资源数据库字段中，以最小公约数的方式为话语资源平台对接提供数据接口。以场景为例，核心元数据标准包括图 2 所示八大类字段。以此为基础，在资源组织过程中，可将对场景的描述做进一步扩充。

在扩展性方面，话语元数据在都柏林核心基础上可以较好的与其它元数据标准耦合扩展，这一点在数据库整合工作中尤为重要，可从数据库结构建模层面减少系统整合的工作量和风险，也为后续与其它资源和前期研究基础相融合奠定了基础。

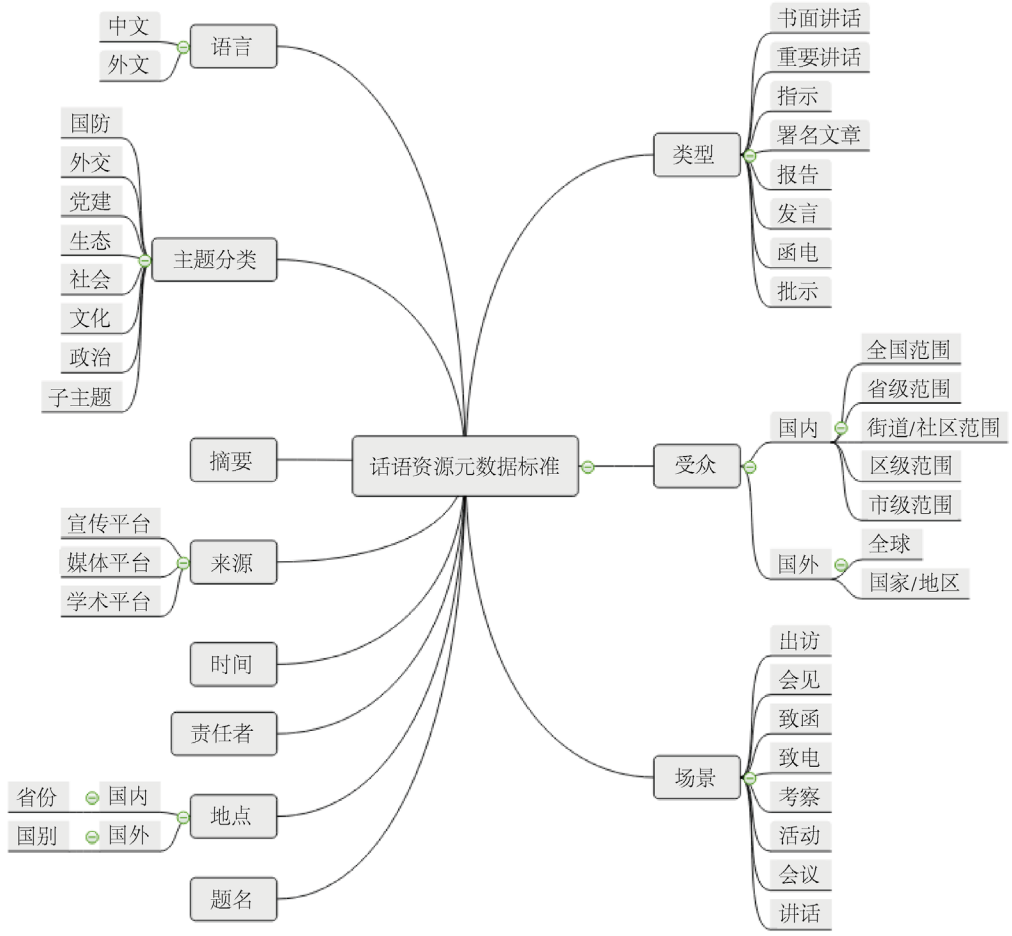

Figure 2. The core standard of discourse resource metadata
图 2. 话语资源元数据核心标准

在抽象性方面，话语元数据核心标准沿用的都柏林标准的设计思想，将抽象性置于标准的核心地位，为更好的与现有系统和其它数据库资源对接，立足于当下宣传思想政治工作的资源现状，同时面向此类资源未来的发展趋势，话语资源元数据核心标准将数据抽象为 11 个大类和若干个可扩展的元素集合。

在丰富性方面，话语元数据核心标准囊括了当下宣传思想政治工作和话语研究的各类场景中的不同维度和类型。面向习近平新时代中国特色社会主义思想，话语元数据可以充分满足该领域资源的组织和管理要求。

6. 结语

话语元数据核心标准既符合都柏林元数据标准的实用性和简单性，也满足了以“习近平思想”话语为代表的中国特色社会主义理论话语资源科学组织的现实需求，融入了时空场景的维度，为后续针对话语资源开展研究奠定了基础。未来随着 5G 技术的普及和媒体融合发展，基于话语元数据建立的数据库平台将有力地推动话语的传播和宣传。

基金项目

江苏省社科基金青年项目“基于时空场景的习近平思想文本数据库构建与应用研究”(17TQC004)；江苏警官学院高层次引进人才科研启动项目(JSPIGKZ)。

参考文献

[1] 戴鸿昊, 史建云. 基于语义标注的数据库元数据质量评估方法[J]. 计算机产品与流通, 2020(11): 178.

[2] 张力元, 王军. 古籍数据库分面分类体系设计[J]. 图书馆建设: 1-9[2020-10-03]. https://cc0eb1c56d2d940cf2d0186445b0c858elksslcnki.i.nuaa.edu.cn:4443/kcms/detail/23.1331.G2.20200820.1048.002.html

[3] 黄琪, 曾建勋, 刘伟. 科技资源关联聚合中的元数据框架研究[J]. 中国科技资源导刊, 2020, 52(4): 38-46.

[4] 杨光. 习近平系列重要讲话 大型网络数据库上线[J]. 计算机与网络, 2014, 40(19): 8.

[5] 孙羽佳. 基于“学习强国”平台的全媒体时代主题报道的创新实践[J]. 长江丛刊, 2020(27): 23+40.

[6] 蔡梦玲. 基于 OAIS 的音视频数据库分层元数据模型[J]. 图书馆杂志, 2019, 38(1): 24-29+35.