

## 文章

[Jingwei Wang](#) · 十月 28, 2021 阅读大约需 32 分钟

# IRIS 2021 技术文档 First Look 31 : InterSystems IRIS 和 UIMA

[\\*/ 目录 技术概要 : InterSystems IRIS 和 UIMA关于 UIMAInterSystems IRIS 如何补充 UIMA Annotation Store \(注释存储\) InterSystems IRIS NLP在 InterSystems IRIS 中体验 UIMA 用前须知 创建新的命名空间 运行Java Gateway \(Java 网关\) 体验 UIMA 分析管道 浏览 Annotation Store \(注释存储\) 通过分析管道发送新文本 向 Aviation.Event 表添加记录 查看Annotation Store \(注释存储\) 中的新条目了解更多关于 UIMA的资料](#)

## 技术概要 : InterSystems IRIS 和 UIMA

本技术概要 (First Look) 快速介绍了 InterSystems IRIS® 数据平台如何实施和补充非结构化信息管理架构 (UIMA)。在简要概述 UIMA 以及 InterSystems IRIS 如何对其进行补充之后, 您将有机会通过基本的动手练习来在实际应用中了解 InterSystems IRIS。要浏览技术概要 (First Look) 的所有内容, 包括可以在InterSystems IRIS的免费评估实例上执行的内容, 请参阅InterSystems 技术概要。

## 关于 UIMA

UIMA 是管理分析非结构化信息 (如文本和视频) 的标准。对于非结构化信息, 计算机通常需要几个步骤才能将信息转化为可操作的结构化数据。例如, 扫描的文档将文本转变为机器可读版本之前需要OCR, 即使这样, 在应用额外的 NLP 策略之前, 计算机也不能很好地处理自然语言文本。由于像这样的流程包括本质上非常不同的步骤, 因此单个工具不太可能处理所有这些步骤。更可能的情况是, 这个流程包括单独模块, 这些模块由使用不同技术的不同主体执行, 且需要协同工作。在 UIMA 中, 这些模块被称为analysis engines (分析引擎)。因为符合 UIMA 标准的分析引擎都遵循相同的标准, 所以它们可以组合成一系列分析器 (a UIMA analysis pipeline [UIMA 分析管道]), 每个部分做自己最擅长的事情。源非结构化数据在通过此 UIMA 分析管道时不会发生改变, 而是沿途生成annotations (注释)。UIMA 标准确保来自一个分析引擎的注释不会干扰来自不同分析引擎的注释。对于文本, 这些注释基于文本中的字符位置。UIMA 的互操作性允许您将来自不同供应商和技术的分析引擎组合到一个管道中, 而无需编写任何自定义代码, 并且由于分析引擎引用原始源数据中的字符位置, 因此它们的注释可以组合、比较和推理。UIMA 标准包括运行这些分析引擎中所执行的框架。除了提供互操作性之外, UIMA 还提供了用于扩展和部署这些分析引擎的框架。这使供应商可以专注于开发分析引擎, 而不必担心扩展和部署他们的解决方案。[How InterSystemsIRISComplementsUIMA .anchor]UIMA 标准还提供了在分布式架构中调用这些分析引擎的框架。每个符合 UIMA 的分析引擎都必须附带一个 XML 描述符文件, 该文件包含基本识别信息, 例如分析引擎的名称和供应商。它还定义了对分析引擎生成的注释进行分类的注释类型。

## InterSystems IRIS 如何补充 UIMA

InterSystems IRIS 在三个方面补充了 UIMA :

- 引入功能索引以创建 UIMA 分析管道 (analysis pipeline), 并在 InterSystems IRIS 表中插入或更新记录时自动为该分析管道提供新文本。
- 将 UIMA 分析管道生成的注释存储在 InterSystems IRIS 中清晰、可通过 SQL 访问的Annotation Store (注释存储) 中。

在 InterSystems IRIS 中体验 UIMA

- 确保 InterSystems IRIS Natural Language Processing (InterSystems IRIS 自然语言处理, NLP) 符合 UIMA 标准, 并可用作 UIMA 分析管道中的分析引擎。

## 创建和调用 UIMA 分析管道

InterSystems IRIS 使用 functional index (功能索引) 以创建 UIMA 分析管道, 该管道使用 InterSystems IRIS 概念, 而无需担心实现 Java 接口。功能索引是 InterSystems IRIS 数据库的一项功能, 它允许在表中插入或更新记录时执行功能。在这种情况下, 功能索引是定义在表的列上, 该列包含您希望由 UIMA 分析管道分析的非结构化数据。设置该管道就像将分析引擎描述文件的位置添加到功能索引定义一样简单。

一旦定义了功能索引, 只要在索引表列中插入或更新最新数据, InterSystems IRIS 就会自动将非结构化数据输入 UIMA 分析管道。例如, 如果在包含报表的列上定义了功能索引, 那么将新报表添加到表中后立即对其进行分析。没有 `InterSystems IRIS` 中的这一特殊功能, 每次您想要分析数据时, 都需要在 Java 中以编程方式通过管道发送非结构化数据。

## Annotation Store (注释存储)

默认情况下, UIMA 分析管道的结果被存储在冗长繁琐的 XML 文件中。由于 UIMA 标准没有提供更复杂的注释存储方法, InterSystems IRIS 通过使用灵活的、基于 SQL 的存储, 将注释放在统一、持久的表中以便后续检索, 从而扩展了 UIMA 分析管道。这种存储系统称为 Annotation Store (注释存储)。

在您第一次编译包含为创建 UIMA 分析管道定义的功能索引的类时, 会自动创建此 Annotation Store (注释存储)。它直接链接到包含非结构化数据的原始表的列。

在架构上, Annotation Store (注释存储) 是通过添加一个特殊的分析引擎作为 UIMA 分析管道的最后一个组件来生成的。当您向 UIMA 功能索引添加到 InterSystems IRIS 类时, 将会自动生成 Annotation Store (注释存储)。在 InterSystems IRIS 之外开发的 UIMA 分析管道也可以将此特殊分析引擎添加到管道末端, 以创建 Annotation Store (注释存储)。这种实现方式超出了本技术概要 (First Look) 的讨论范围。

您还可以使用 XData block 在包含功能索引的类中自定义 Annotation Store (注释存储)。例如, 您可以为每个表定义额外的列和索引。您还可以筛选注释类型, 以将它们排除在 Annotation Store (注释存储) 之外。

## InterSystems IRIS NLP

InterSystems IRIS Natural Language Processing (InterSystems IRIS 自然语言处理, NLP) 被内置到 InterSystems IRIS® 数据平台中, 可以在预先不了解这些非结构化文本主题的情况下, 对这些文本进行文本分析。它通过应用识别语义实体的特定语言规则来实现这项功能。由于这些规则是专门针对语言而非内容, 因此 InterSystems IRIS NLP 可以在不使用字典或本体的情况下提供对文本内容的深入分析

您可以使用 InterSystems IRIS NLP 作为 UIMA 分析引擎, 为 NLP 概念和上下文生成 UIMA 注释。这些注释与其他 UIMA 分析引擎提供的 UIMA 注释完全兼容。

## 在 InterSystems IRIS 中体验 UIMA

现在您已经了解了有关 UIMA 的一些基本信息, 可以动手体验以了解它在 InterSystems IRIS 中的工作原理。在体验之前, 您需要设置环境。

在 InterSystems IRIS 中体验 UIMA

## 用前须知

首先，请执行以下初步设置任务：

0. 安装 Java Runtime Environment (Java 运行时环境)。
1. 安装 InterSystems IRIS。
2. 创建一个新的 InterSystems IRIS 命名空间。
3. 启动 Java Gateway (Java 网关)。

### 安装 Java Runtime Environment (Java 运行环境)

InterSystems IRIS 进行 UIMA 分析管道分析需要安装 Java Runtime Environment (Java 运行时环境, JRE)。还需要一个指向 JRE 安装位置的环境变量。

0. 如果您的机器上尚未安装 JRE，请从 Oracle® 下载并安装最新版本。
1. 创建一个名为 JAVA\_HOME 的环境变量，该变量指向 JRE 安装的位置。例如，在 Windows® 上，使用 Control Panel (控制面板) 创建 JAVA\_HOME 环境变量，并定义该变量到 JRE 安装位置的路径。

### 安装 InterSystems IRIS

要运行 UIMA 分析管道的演示，您需要一个运行中的 InterSystems IRIS 已授权的实例。

有关如何安装和授权 InterSystems IRIS 开发实例的说明，请参阅 InterSystems IRIS Basics: Installation (《InterSystems IRIS 基础知识：安装》)。

## 创建新的命名空间

作为技术概要 (First Look) 体验的一部分，您将向 InterSystems IRIS 中的命名空间添加一个新类文件。为了使此示例数据与预定义的命名空间分开，请创建一个名为 SAMPLES 的新命名空间来保存与此技术概要 (First Look) 关联的代码和数据。创建一个新的命名空间：

0. 在浏览器中打开 Management Portal (管理门户)，使用 InterSystems IRIS Basics: Connecting an IDE (《INTERSYSTEMS IRIS 基础：连接一个 IDE》) 的实例适用的 URL。
1. 选择 System Administration (系统管理) > Configuration (配置) > System Configuration (系统配置) > Namespaces (命名空间)。
2. 在 Namespaces (命名空间) 页面，选择 Create New Namespace (新建命名空间)。
3. 在 New Namespace (新建命名空间) 页面，输入 SAMPLES 作为新命名空间的名称。
4. 点击 Create New Database (新建数据库)，将显示 Database Wizard (数据库向导)。
5. 在 Database Wizard (数据库向导) 的第一页，在 Enter the name of your database (输入您的数据库名称) 字段，输入您正在创建的数据库的名称，例如 Samplesdb。
6. 输入数据库的目录，例如 C:\InterSystems\IRIS\mgr\Samplesdb。
7. 点击 Next (下一步)。
8. 点击 Finish (完成)。
9. 回到 New Namespace (新命名空间) 页面，在 Select an existing database for Routines (为 Routine 选择一个现有数据库) 下拉菜单，选择您刚刚创建的数据库 Samplesdb。
10. 点击靠近页面顶部的 Save (保存)，然后在结果日志的末尾单击 Close (关闭)。

UIMA 分析管道

## 运行 Java Gateway (Java 网关)

Java Gateway (Java 网关) 可以实例化外部 Java object (Java 对象), 就好像它是 InterSystems IRIS 中的本机对象一样, 可以对其进行操作。InterSystems IRIS 的 UIMA 策略使用 Java Gateway (Java 网关), 可以从 command line (命令行) 启动 Java Gateway (Java 网关)。例如, 在 Windows 上:

0. 打开 Run (运行) 对话框。

1. 输入以下命令:

```
%JAVAHOME% bin java
-classpath "C:\InterSystems\IRIS\dev\java\lib\JDK18\*;C:\InterSystems\IRIS\dev\java\lib\jackson\*;C:\InterSystems\IRIS\dev\java\lib\luma\*" com.intersystems.gateway.JavaGateway 5555
```

其中:

- JAVAHOME 是指向 Java Runtime Environment (Java 运行环境, JRE) 安装目录位置的环境变量。
- C:\InterSystems\IRIS 是您安装的 InterSystems IRIS 的目录。
- JDK18 对应您的 JRE 版本。

如果在 UNIX® 上运行, 请记住 -classpath 的句法使用冒号作为分隔符。

## 体验 UIMA 分析管道

现在您已经完成了准备工作, 并已准备好在实际应用中查看 UIMA 分析管道的运行情况。在这次体验中, 您将:

- 添加包含 UIMA 功能索引的类文件
- 编译包含功能索引的类。
- 查看 Annotation Store (注释存储) 的表。
- 将非结构化数据添加到示例数据库。
- 浏览 Annotation Store (注释存储) 以获取分析管道生成的新数据。

### 添加带有 UIMA 功能索引的类文件

您可以通过为包含非结构化文本的表定义功能索引, 向 UIMA 分析管道添加分析引擎。在本次体验中, 您将向该分析管道添加 InterSystems IRIS NLP 分析引擎。

在体验的这一部分中, 您将创建一个新的类文件。可以在您喜欢的文本编辑器或 IDE (例如 InterSystems Studio 或带有 ObjectScript 扩展名的 Visual Studio Code) 中创建类文件。

0. 在 IDE 或文本编辑器中创建一个新文件。

1. 将以下内容复制并粘贴到类文件中:

```
XXXXXXXXXX
```

```
Class Sample.MyData Extends %Persistent
```

```
?
```

```
{
```

```
?
```

```
Property MyText As %String;
```

```
?
```

```
Index MyIndex On (MyText) As %UIMA.Index(AEDESCRIPTOR = \  
"classpath:/com/intersystems/uima/annotator/iKnowEngine.xml");
```

```
?
```

```
}
```

体验 UIMA 分析管道

其中：

- MyText 是包含非结构化文本的 Sample.MyData 表中的列。
- MyIndex 是 UIMA 功能索引。
- iKnowEngine.xml是 InterSystems IRIS NLP 分析引擎的描述符文件。

3. 将文件另存为sample.cls。

## 编译表类

要自动生成 Annotation Store（注释存储），您只需使用 InterSystems Terminal（InterSystems 终端）或 IDE 提供的工具编译包含功能索引的类。

**建议：**使用 InterSystems Terminal（InterSystems 终端）时，您可以使用 Shift+Insert 命令将剪贴板的内容粘贴到终端命令提示符中。从本指南复制命令并将它们粘贴到终端中以减少错误，这种方式很有用。

加载和编译类：

0. 打开 InterSystems Terminal（InterSystems 终端）。有关为您的实例打开 Terminal（终端）的信息，请参阅 InterSystems IRIS Basics: Connecting an IDE（《INTERSYSTEMS IRIS 基础：连接一个 IDE》）中的 InterSystems IRIS Connection Information（《InterSystems IRIS 连接信息》）。

1. 切换到您为此演示创建的命名空间。例如：`set $namespace="samples"`

2. 输入以下命令，将类文件加载到命名空间中：`do $system.OBJ.Load("<sample-dir>/sample.cls")`

其中 <sample-dir> 是您保存 samples.cls 类文件的位置。

4. 输入以下命令，编译您粘贴到 sample.cls 的 Sample.MyData 类：`do $system.OBJ.Compile("Sample.MyData")`

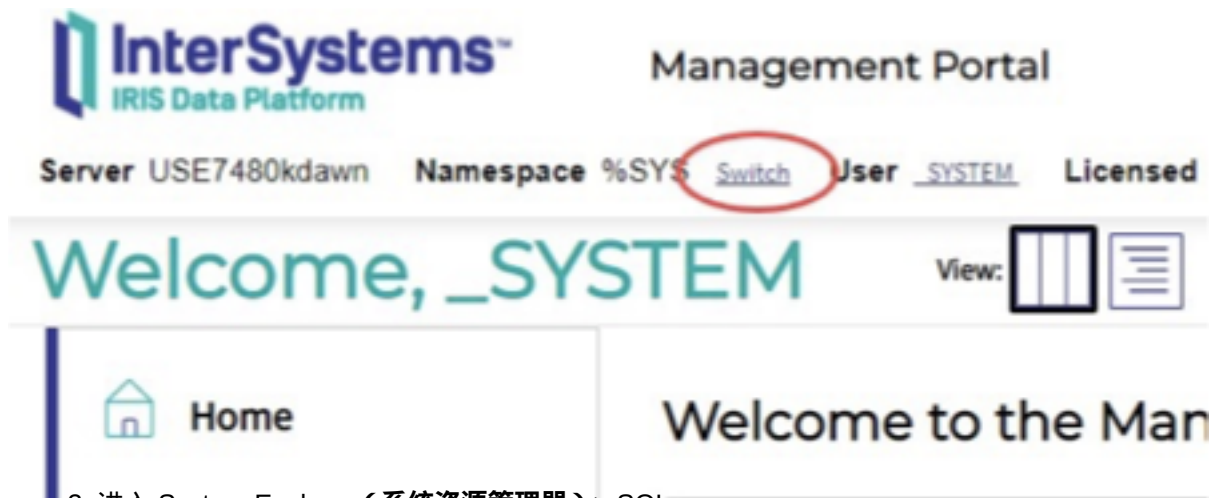
## 浏览 Annotation Store（注释存储）

既然您已经使用 Annotation Store UIMA 功能索引编译了类，可以浏览为保留 InterSystems IRIS NLP 生成的注释而创建的（Annotation Store）注释存储。

0. 在浏览器中使用 URL 为实例打开 Management Portal（管理门户），如 InterSystems IRIS Basics: Connecting an IDE（《INTERSYSTEMS IRIS 基础：连接一个 IDE》）的 InterSystems IRIS Connection Information（InterSystems IRIS 连接信息）所述。

1. 使用标题中的链接切换到 Samples 命名空间。

参观 UIMA 分析管道



3. 进入 System Explorer（系统资源管理器）> SQL。

4. 展开左侧窗格中的 Tables（表）列表。

可以看到 Annotation Store（注释存储）的三个表。这些表的命名规则对应包含所分析的非结构化文本的表（Sample.MyData）。

- SampleMyData.Type-----包含此存储中使用的 Annotation Types（注释类型）的概述。
- SampleMyData.Sofa-----包含 sofas，这是 UIMA 分析引擎分析的文本对象。
- SampleMyData.Annotation-----包含该分析引擎生成的注释。

您可以修改功能索引定义以创建多个注释表，然后根据注释类型将结果输出到正确的表中。

## 通过分析管道发送新文本

InterSystems IRIS 中 UIMA

分析管道的强大之处在于，新的非结构化文本会自动通过管道发送以进行分析，并将结果添加到 Annotation Store（注释存储）中。由于您已经创建了 Annotation Store（注释存储），可以看到新纪录如何被添加到 Sample.MyData 表，并生成添加到 Annotation Store 的新条目。

### 向 Aviation.Event 表添加记录

在本步骤中，您将使用 SQL 向示例数据库中的 Sample.MyData 表添加一些非结构化文本。请记住，该表包含 MyText 列，您已在该列上定义功能索引。正如您将看到的，当进行此插入操作时，会自动生成注释。

0. 在 System Explorer（系统资源管理器）> SQL 页，展开左侧窗格中的 Tables（表）列表。
1. 选择 Sample.MyData，这是包含通过分析管道发送的非结构化文本的表。
2. 在右侧窗格中，单击 Execute Query（执行查询）标签。
3. 要将新条目插入示例数据库，请在文本框中输入以下查询

```
XXXXXXXXXX
```

```
INSERT INTO Sample.MyData (MyText) VALUES ('\First Look unstructured text\')
```

5. 点击 Execute（执行）。

这会将短语 "First Look unstructured text" 放入 Sample.MyData 表的 MyText 列中。

了解更多关于 UIMA 的资料

**查看 Annotation Store（注释存储）中的新条目**

现在您已将新的非结构化文本添加到示例数据库中，可以查看 Annotation Store 以了解这些文本是如何通过分析管道自动发送的。您可以看到新的非结构化文本和来自 InterSystems IRIS NLP 的注释都添加到了 Annotation Store（注释存储）中。

0. 在 System Explorer（系统资源管理器）> SQL 页，展开左侧窗格中的 Tables（表）列表。
1. 选择 SampleMyData.Sofa 表。
2. 在右侧窗格中，单击 Open Table（打开表）。



4. 点击 Close Window（关闭窗口）。
5. 在左侧窗格中，选择 SampleMyData.Annotation 表。
6. 点击 Open Table（打开表）。

在 coveredText 列，您可以看到由 InterSystems IRIS NLP 分析引擎生成的注释。

## 了解更多关于 UIMA 的资料

要了解有关 InterSystems IRIS 如何实施和补充 UIMA 的更多信息，请参阅 Using InterSystems UIMA（《使用 InterSystems UIMA》）。

有关 UIMA 标准的框架、基础设施和组件的详细概述，请参阅 [Apache UIMA home page](#)

[#UIMA](#) [#Ensemble](#) [#InterSystems IRIS](#) [#InterSystems IRIS for Health](#)

### 源

URL:

<https://cn.community.intersystems.com/post/iris-2021-%E6%8A%80%E6%9C%AF%E6%96%87%E6%A1%A3-first-look-31%E5%92%8C-uima>