Grupos de trabajo  RTIRIS-5

Reunión sobre Sistemas de Búsqueda en la Red

Índice

- Directorio X.500 y LDAP
  - Cambios en el directorio para el soporte de búsquedas integradas mediante LDAP
  - Otras opciones
  - Herramientas
- Grupo sobre indexación iris-index
  - Piloto iris-index
  - Otras opciones
  - Herramientas

Directorio X.500 y LDAP

- Cambios en el Directorio para el soporte de búsquedas integradas mediante LDAP
  - Objetivo:
    - Un usuario necesita la dirección de correo electrónico de una persona desde su programa de correo
  - Ejemplo:
    - Desde Netscape

Búsquedas integradas mediante LDAP

- Actualidad
  - Búsqueda en 1 organización

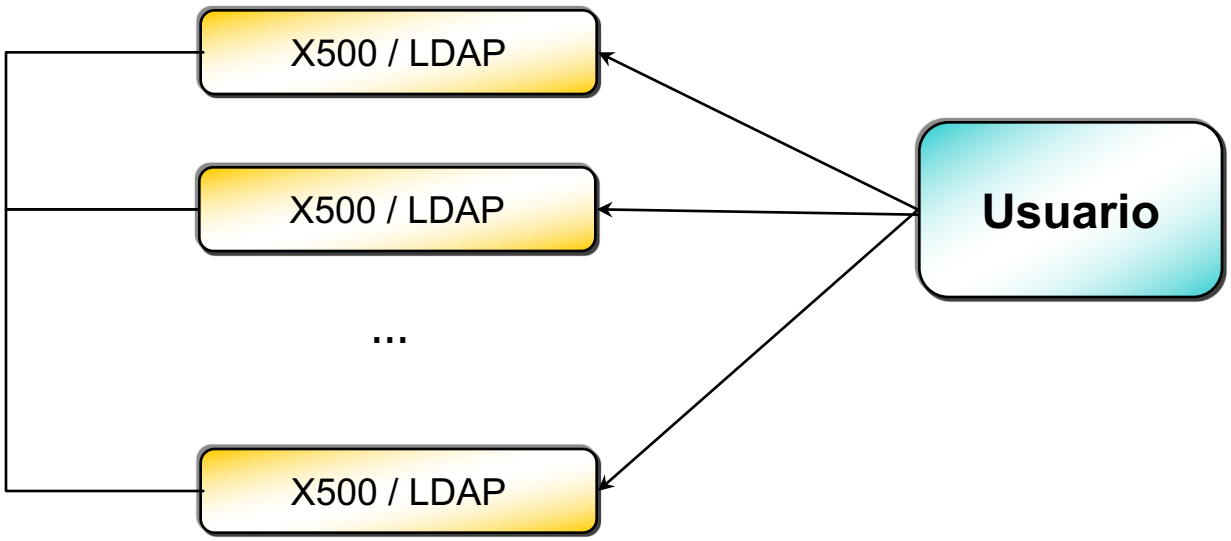
Name:	Prueba LDAP
LDAP Server:	ldap.xxx.es
Search Root:	o=xx,c=es

- Futuro
  - Búsqueda en múltiples organizaciones

Name:	Prueba LDAP
LDAP Server:	ldap.xxx.es
Search Root:	c=es

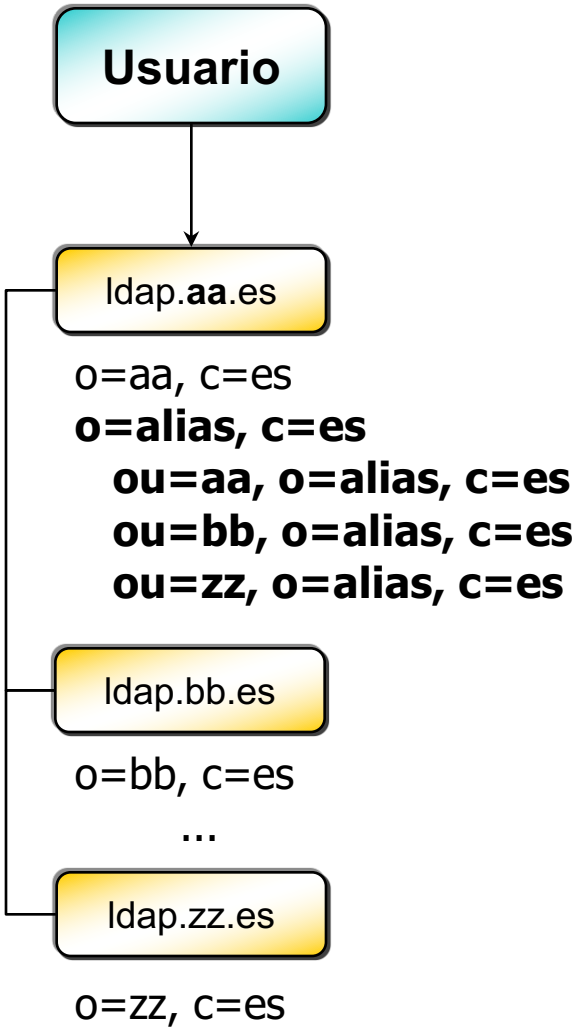
¿ Cómo lo hacemos ?

Opción 1: Búsquedas en el Cliente



- El cliente realiza las búsquedas en cada uno de los servidores que existen en ese nivel
- Se encarga de ordenar las respuestas
- Cuenta *directorio*

Opción 2: Búsquedas mediante alias



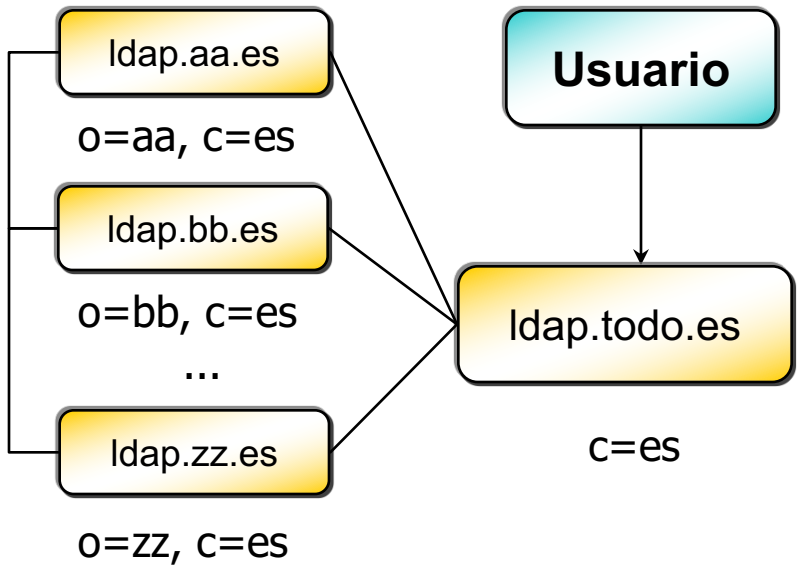
Name:	Prueba LDAP
LDAP Server:	ldap.aa.es
Search Root:	o=alias,c=es

- Se crea **o=alias** con alias a las demás organizaciones de c=ES
- La búsqueda se hace en **ldap.aa.es** y éste consulta a los demás DSAs
- **ldap.aa.es** espera a que se devuelvan las respuestas
- Se encarga de ordenar todas las respuestas

Opción 2: Búsquedas mediante alias

- Un servidor tendría que replicar la pregunta a todos los servidores de un mismo nivel en la estructura jerárquica
  - Cada servidor ha de procesar la pregunta y buscar los datos incluso aunque no los tenga
  - Carga excesiva de las máquinas
- Por esto no se permiten hacer búsquedas a nivel del mundo ni a nivel de un país entero

Opción 3: Búsquedas en un servidor LDAP duplicado



Name:	Prueba LDAP
LDAP Server:	ldap.todo.es
Search Root:	c=es

- Tenemos un servidor LDAP con réplicas de lo que hay en c=es
- Los usuarios realizan búsquedas en ese servidor
- Se replica:
  - Nombre,
  - Apellidos,
  - Email,
  - Teléfono,
  - Fax
- Se incorpora:
  - DN de la entrada original para obtener el resto de atributos
  - Fecha de la última actualización

Opción 3: Búsquedas en un servidor LDAP duplicado

- Problemas:
  - ¿ Cómo actualizamos la información ?
    - Volcados masivos
    - Programa que interroga la fuente y el destino y vuelca modificaciones
  - ¿ Cuándo la actualizamos ?
    - Periódicamente por la noche
    - Cuando exista una modificación en una fuente
  - ¿ Quién tiene permiso para hacerlo ?
    - Los *Directory Manager*s de cada organización

Otras Opciones

- Servicio piloto NameFLOW LDAP
- Servidor Whois++ y
  red de servidores Whois++

Servicio Piloto NameFLOW LDAP

- Origen:
  - Problemas con X.500 (93)
    - Versiones del sofware disponibles
    - Dos test de pilotos infructuosos
  - Problemas con software Quipu para fechas posteriores al año 2.000
  - Se hace necesaria la migración de la infraestructura X.500(88) a una basada en el protocolo LDAP
  - Especialmente a LDAPv3
    - Permite realizar *referrals*
    - Almacena los esquemas de las entradas en el propio directorio

Servicio Piloto NameFLOW LDAP

- Objetivos:
  - Mover hacia una arquitectura basada en productos baratos, abiertos, fáciles de manejar, fáciles de ampliar, ...
  - Proporcionar índices/centroides de la información contenida en el directorio (para agilizar las búsquedas)
  - Proporcionar compatibilidad con versiones Quipu (88) si fuese necesario

Servicio Piloto NameFLOW LDAP

- Necesitaremos
  - Robots LDAP que sean capaces de recolectar la información de los índices generados en cada servidor LDAP
- Servicios a obtener
  - Infraestructura de conexión de servidores LDAP
  - Mecanismos para intercambiar la información de índices/centroides LDAP
  - Realizar búsquedas usando índices y/o centroides

Red de servidores Whois++

- Servidor Whois++
  - Basado en una serie de plantillas estandarizadas de conjuntos ordenados de pares atributo-valor
  - Cada registro está estructurado dentro de una plantilla y tiene un identificador único
  - Cada servidor Whois++ necesita un identificador único en Internet
    - Ha de registrarse en una organización que da el identificador
- Búsquedas simples mediante comandos:
  - Término/os de búsqueda + restricciones opcionales
    - Tipo de plantilla
    - Valor de un atributo
    - Identificador de registro

Red de servidores Whois++

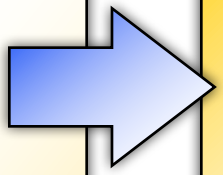
- Arquitectura
  - Servidor base. Contiene las plantilla rellenas
  - Servidor de índices que contiene
    - Base de conocimiento
    - Punteros hacia otros servidores de índices o a servidores de base
- Centroide
  - Es un tipo de base de conocimiento
  - Lista de plantillas y atributos usados por el servidor +
  - Lista de palabras para cada atributo
    - Contiene una ocurrencia de cada una de las palabras que aparece al menos una vez en ese atributo en algún registro de los datos de ese servidor

Centroide - Ejemplo

```
Registro 1
Template:         Person
First-Name:       Javi
Last-Name:        Massa
Favorite-Drink:   Agua

Registro 2
Template:         Person
First-Name:       Celestino
Last-Name:        Tomás
Favorite-Drink:   Cerveza

Registro 3
Template:         Domain
Domain-Name:      rediris.es
Contact-Name:     Pepe López
```



Centroide

```
Template:         Person
First-Name:       Celestino
                  Javi
Last-Name:        Massa
                  Tomás
Favorite-Drink:   Agua
                  Cerveza

Template:         Domain
Domain-Name:      rediris.es
Contact-Name:     Pepe
                  López
```

Indexado en Whois++

- Sirve para agrupar servidores de base y de indexado en un servicio de directorio unificado
- Cada servidor ha de generar una base de conocimiento para las entradas que contiene en sus bases de datos
- Un tipo de base de conocimiento se consigue construyendo un centroide
- Los servidores de índices pueden recolectar estas bases de conocimiento de otros servidores

Preguntas a un servidor de índices Whois++

- Un servidor de índices puede:
  - Tomar una pregunta
  - Buscar:
    - En su colección de centroides
    - Otra base de conocimiento que tenga
  - Determinar los servidores que mantienen registros que pueden coincidir con la pregunta
  - Notificar al cliente los siguientes servidores con los que contactar para enviarles la pregunta
- Control de bucles
  - Numeramos cada nivel de indexación
  - El cliente gestiona una lista de servidores a los que ha de preguntar

Herramientas y Software

- Disponibles
  - Net_LDAPapi-1.40
    - Librería de acceso desde Perl5 a la API creada por la Universidad de Michigan y Netscape
    - Posee una pasarela Web a X500 como ejemplo
  - Web500gw-2.1b2
    - Contacto con Frank Richter para colaboración
- Necesarias
  - Estadísticas de acceso por web500gw
  - Control de búsquedas de robots
    - Crear cgi que sea el que llame a web500gw
    - Frank Ritcher devuelve **robot.txt** a petición del robot
    - Modificaciones para inclusión de <META NAME=robot ...>

Software

- Digger 2.0. Servidor Whois++
- Software IC-R4.0
  - Disponible

- Estadísticas de accesibilidad (estadoDSAs)
  - Mayoría por encima del 80 %
  - Varios DSAs por debajo del 5%

Grupo sobre indexación iris-index

- Metainformación
- Piloto indexación con información válida
- Otras posibilidades
- Herramientas

Metainformación

- Uso de metainformación basado en el conjunto de elementos de Dublin Core
  - DC es una iniciativa internacional que define un conjunto de elementos para la descripción de recursos
    - 15 elementos
    - Admiten cualificadores

Metainformación - Formato a usar

Posibilidades

1. `<META NAME="DC.Creator" CONTENT="(TYPE=name) Javier Massa">`
   `<META NAME="DC.Creator" CONTENT="(TYPE=email) prueba@rediris.es">`

2. `<META NAME="DC.Creator.name" CONTENT="Javier Massa">`
   `<META NAME="DC.Creator.email" CONTENT="prueba@rediris.es">`

3. `<META NAME="DC.Creator" TYPE="name" CONTENT="Javier Massa">`
   `<META NAME="DC.Creator" TYPE="email" CONTENT="prueba@rediris.es">`

SOIF generados con Harvest

1. `dc.creator{55}: (TYPE=name) Javier Massa`
   `(TYPE=email) prueba@rediris.es`

2. `dc.creator.email{17}: prueba@rediris.es`
   `dc.creator.name{12}: Javier Massa`

3. `dc.creator{30}: Javier Massa`
   `prueba@rediris.es`

Metainformación - Elementos <META>

- Hemos tomado el caso tercero con los siguientes elementos
  - NAME
    - Nombre del elemento
  - TYPE
    - Permite un refinamiento y la clarificación del contenido del elemento
  - SCHEME
    - Nos indica el contexto en el que hemos de interpretar el valor del elemento
  - CONTENT
    - Valor del elemento

Metainformación - Cambios

- En la última reunión de DC se ha cambiado el número de elementos que componen <META>
  - NAME
  - SCHEME
  - LANG
    - Especifica el lenguaje en el que está el valor de un elemento
  - CONTENT
- Se renombra TYPE a SUBELEMENT para evitar confusiones

Metainformación - Inclusión en HTML

- Se han diseñado varios mecanismos para la inclusión de los cualificadores en HTML
  - HTML 3.2

```
<META NAME="DC.Creator" CONTENT="(TYPE=name) Javier Massa">
<META NAME="DC.Creator" CONTENT="(TYPE=email) prueba@rediris.es">
```

  - HTML 4.0

```
<META NAME="DC.Creator.Name"  CONTENT="Javier Massa">
<META NAME="DC.Creator.Email" CONTENT="prueba@rediris.es">
```

- Hay un grupo de trabajo trabajando en SUBELEMENT para incorporar los cualificadores necesarios

Metainformación - Cualificadores

- Title
  - Alternative
  - Main
- Creator
  - PersonalName
    - Address
  - CorporateName
    - Address
- Publisher
  - PersonalName
    - Address
  - CorporateName
    - Address
- Contributor
  - PersonalName
    - Address
  - CorporateName
    - Address
- DATE
  - Created
  - Issued
  - Accepted
  - Avaliable
  - Acquired
  - DataGathered
  - Valid
- Relation
  - Type
- Coverage
  - PeriodName
  - PlaceName
  - x
  - y
  - z
  - t
  - Polygon
  - Line
  - 3d

Metainformación - Herramientas

- Necesidad de actualizar nuestro formato a los nuevos cambios de DC
  - Se esperan más cambios
- Modificación de MetaWebber
  - para adaptarlo a este formato de DC
  - para adaptarlo a futuras modificaciones de DC
- Uso de editores de metainformación
  - dcdot.pl
    - Dado un URL lo chequea y edita un formulario con los metas que contiene y nos dejar añadir otros
  - Herramientas hechas en RedIRIS para incorporación interactiva en ficheros (uso con CVU)

Piloto de indexación con información válida

- Se indexarán
  - Registro de recursos
    - Listas de distribución
    - Servidores web
    - Bibiliotecas
  - Piloto iris-index
    - Metainformación en formato DC
    - Palabras entre `<H>` y `</H*>`
    - Palabras entre `<TITLE>` y `</TITLE>`
  - Comunidades Virtuales de Usuarios (CVUs)
  - ¿ Proyecto DisEven ?
  - ¿ Directorio LDAP/X.500 ?

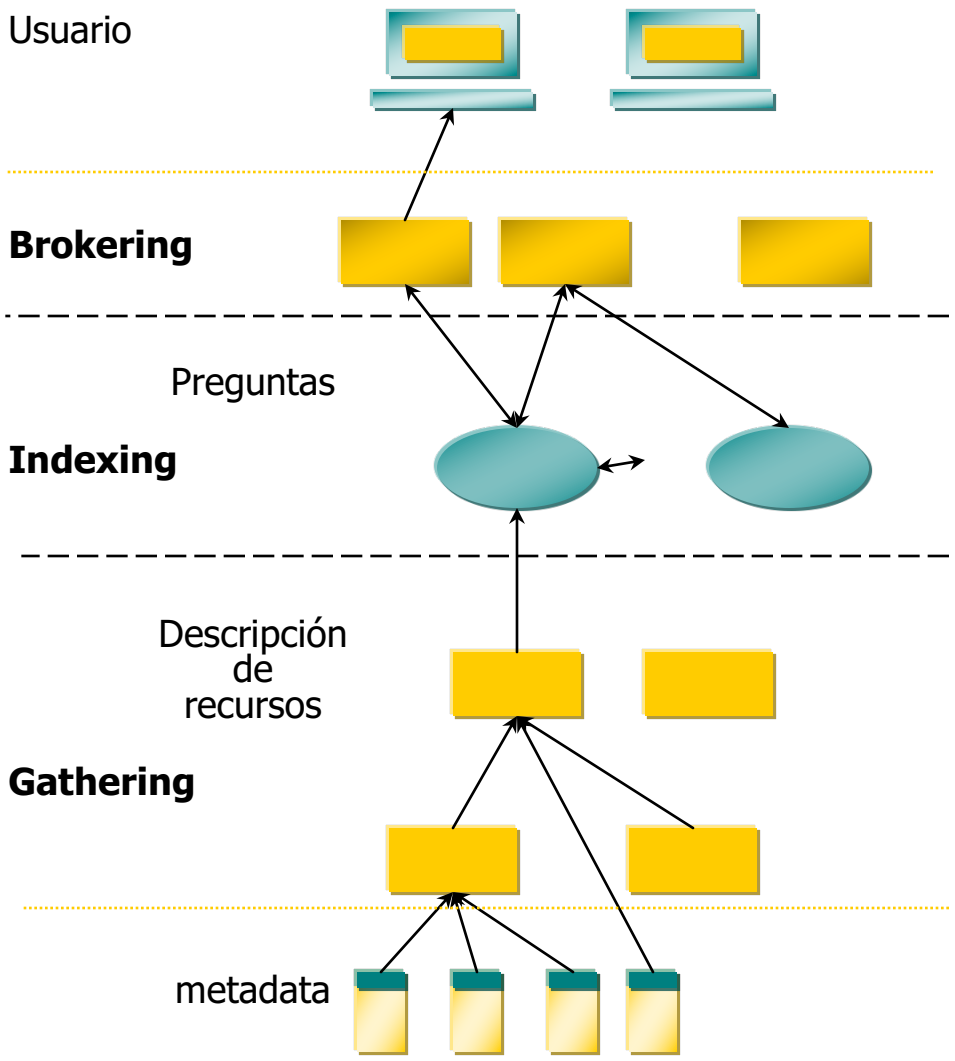
Piloto de indexación iris-index

- Se utilizarán herramientas para incorporar metainformación de forma cómoda
- Se usará Harvest 1.5 para este piloto
- Cada centro participante indexará la información de sus servidores
- RedIRIS tendrá varios *brokers* para poder buscar en diferentes secciones del índice
- Mostrará los resultados usando los valores contenidos en la metainformación

Otras posibilidades de indexación

- Piloto CHIC
  - Se necesita encontrar información de la comunidad investigadora y académica
  - Desarrollar y probar una arquitectura de indexación para nuestra comunidad que
    - Permita crecer para albergar una gran cantidad de información
    - Proporcione calidad en las búsquedas usando exclusivamente metainformación
  - Seleccionar estandares que permitan su uso con software de dominio público y comercial
    - SOIF para la descripción de recursos
    - DC para el formato de la metainformación

Piloto CHIC - Arquitectura



- Interface de búsquedas a los usuarios
- *Routing* de preguntas por la red de servidores indexados
- Transformar las preguntas en lenguaje original al lenguaje de la base de datos
- Mezclar los resultados de múltiples fuentes

- Construcción de BD de un conjunto de descriptores de recursos (SOIF)
- Proporcionar un interface de consultas al broker
- Proporcionar pistas sobre el contenido de la BD para los *routings* de consultas

- Recolección de datos y metadatos
- Transformación de éstos en registros de descripción de recursos (SOIF)
- Ofrecer los registros de descripción de recursos a los servidores de índices o a los nodos que recopilan registros de descripción de recursos

Piloto CHIC - Arquitectura

- Interface de búsquedas a los usuarios
- *Routing* de preguntas por la red de servidores indexados
- Transformar las preguntas en lenguaje original al lenguaje de la base de datos
- Mezclar los resultados de múltiples fuentes

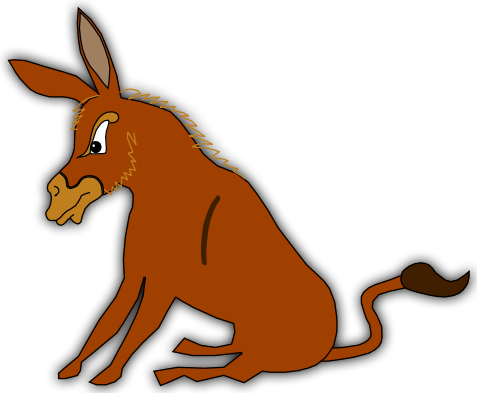
- Construcción de BD de un conjunto de descriptores de recursos (SOIF)
- Proporcionar un interface de consultas al broker
- Proporcionar pistas sobre el contenido de la BD para los *routings* de consultas

- Recolección de datos y metadatos
- Transformación de éstos en registros de descripción de recursos (SOIF)
- Ofrecer los registros de descripción de recursos a los servidores de índices o a los nodos que recopilan registros de descripción de recursos

Herramientas

- Desarrolladas en iris-index
  - Metawebber
    - Incorporación de metainformación a páginas diseñadas
  - Herramientas para CVU
    - Volcado de ficheros a un servidor web desde netscape controlando la incorporación de metainformación
  - Brokerstats - Hermann Straus, dit, upm
    - Estadísticas de consulta a los brokers
- Otras
  - Robot watcher
    - Estadísticas de tráfico producido por los robots en nuestro servidor

Otros Temas



Direcciones de interés

- Más cosas que comentarnos ...
  - **x500**@rediris.es
- Para estar en contacto entre todos
  - **iris-x500**@listserv.rediris.es
  - **iris-index**@listserv.rediris.es