**IPPH国外生物序列数据加工**

# 数据情况

## 20160815V1.7

数据范围：19160321-20160225

|  |  |  |
| --- | --- | --- |
| 概览 | 原始FASTA记录 | 43439799 |
| 原始-EMBL记录 | 43439727 |
| 原始专利号记录 | 254335 |
| 导出专利记录 | 252899 |
| 导出序列记录 | 32831315 |
| 原始-EMBL记录-细览 | 超过30w | 10608392 |
| 未解析专利号 | 19 |
| 未找到标准号 | 1 |
| 正常导出 | 32831315 |
| 原始专利号记录-细览 | 超过30w | 15 |
| 未找到标准号 | 1 |
| 重复专利数 | 1420 |
| 正常导出专利数 | 252899 |
| 正常导出专利国别统计 | country | total |
| WO | 54933 |
| USRE | 1 |
| US | 86044 |
| NO | 1 |
| NL | 4 |
| LU | 3 |
| KR | 19452 |
| JP | 68137 |
| GB | 114 |
| FR | 103 |
| EP | 25335 |
| DE | 173 |
| CH | 3 |
| CA | 7 |
| BE | 6 |
| AU | 15 |
| AT | 4 |
| 正常导出序列数量 | country |  |
| AT | 23 |
| AU | 134 |
| BE | 35 |
| CA | 34 |
| CH | 44 |
| DE | 1962 |
| EP | 5077424 |
| FR | 1123 |
| GB | 1180 |
| JP | 7524086 |
| KR | 857832 |
| LU | 8 |
| NL | 18 |
| NO | 8 |
| US | 12501143 |
| WO | 6866261 |

抽检数据：

localhost:8082/exchange/excute?docids=WO2016010047A1,WO2015190458A1,JP2015534456A,JP2015221760A,JP2016027798A,JP2015536659A,JP2016020336A,JP2016000035A

localhost:8082/exchange/excute?docids=US8101563B2,US6274350B1,US5175099A,US5175099A,US4783524A,US9056899B2,US9068993B2,WO03018792A1,WO2013094240A1,WO2005019832A2,WO2006018211A1,KR920702233A,KR20140015044A,JP2012147703A,JP2009543574A,JP2010508861A,JP1992304888A,JP2004113117A,JP2013521781A,EP2437750A1,EP2522728A1

抽检数据第二次，在修改规则之后

localhost:8082/exchange/excute?docids=JP1992304888A,JP2013511989A,JP2012147703A,JP2014520784A,JP2015230222A,JP2016500257A,JP2005287422A,JP2004135532A,US8309707B2,US6078782A,US5977320A,US6395276B1,US6849417B1,US8614187B2,US8614292B2,US5484906A,US7960106B2,WO9949035A2,WO2016021683A1,WO2015190529A1,WO2016017631A1,WO2007083116A2,WO2013035757A1,WO2006046455A1,WO9749726A1,KR20050085202A,KR19930700853A,KR20120055529A,EP2295584A2,EP1561818A1,EP0153188A1,EP2964718A2

本次进行程序修正，修正履历如下：

针对生物序列st.26版本规范进行调整

主要包括特征关键词、location、特征其他信息等进行规则修改

对应规则文件：国外生物序列索引标签对应说明2016-7-21.xlsx

注：目前关键词提取规则后期仍然需要调整

以下数据记录为旧数据履历情况

## 20160509V1.6

数据范围：19160321-20160211

|  |  |  |
| --- | --- | --- |
| 概览 | 原始FASTA记录 | 43146575 |
| 原始-EMBL记录 | 43146503 |
| 原始专利号记录 | 253618 |
| 导出专利记录 | 252182 |
| 导出序列记录 | 32744723 |
| 原始-EMBL记录-细览 | 超过30w | 10401760 |
| 未解析专利号 | 19 |
| 未找到标准号 | 1 |
| 正常导出 | 32744723 |
| 原始专利号记录-细览 | 超过30w | 15 |
| 未找到标准号 | 1 |
| 重复专利树 | 1420 |
| 正常导出专利数 | 252182 |
| 正常导出专利国别统计 | country | total |
| WO | 54807 |
| USRE | 1 |
| US | 86044 |
| NO | 1 |
| NL | 4 |
| LU | 3 |
| KR | 19452 |
| JP | 67546 |
| GB | 114 |
| FR | 103 |
| EP | 25335 |
| DE | 173 |
| CH | 3 |
| CA | 7 |
| BE | 6 |
| AU | 15 |
| AT | 4 |
| 正常导出序列数量 | country |  |
| AT | 23 |
| AU | 134 |
| BE | 35 |
| CA | 34 |
| CH | 44 |
| DE | 1962 |
| EP | 5077424 |
| FR | 1123 |
| GB | 1180 |
| JP | 7446778 |
| KR | 857832 |
| LU | 8 |
| NL | 18 |
| NO | 8 |
| US | 12501143 |
| WO | 6856977 |

抽检数据：

localhost:8082/exchange/excute?docids=EP2970373A1,EP2917244A1,EP2961419A2,WO2015140313A1,WO2015104564A2,US8829199B2,US8722353B2,US8962267B1,US8940690B2,US8900824B1,US9040264B2,US9040253B2,US9051392B2,US9045547B2

localhost:8082/exchange/excute?docids=US8283330B2,US8273723B2,US8257695B2,US8535725B2,EP2898088A1,EP2863946A2,WO2007136815A2,WO2008032007A1,DE19906169A1,JP2005000071A,JP2004344160A,JP2009539347A,KR20020001873A,FR2632656A1,GB2210883A

localhost:8082/exchange/excute?docids=KR20150059649A,KR20150059449A,KR20150058671A,KR20060015558A,KR20060054176A,KR20060020616A,KR20060003903A,KR20060035601A,KR20060018868A,KR20060030038A,KR20060022283A,KR20060070491A,KR20060111877A

## 20160108V1.5

数据范围：19160321-20151016

|  |  |  |
| --- | --- | --- |
| 概览 | 原始FASTA记录 | 40838110 |
| 原始-EMBL记录 | 40838038 |
| 原始专利号记录 | 245472 |
| 导出专利记录 | 244036 |
| 导出序列记录 | 30726993 |
| 原始-EMBL记录-细览 | 超过30w | 10110583 |
| 未解析专利号 | 19 |
| 未找到标准号 | 443 |
| 正常导出 | 30726993 |
| 原始专利号记录-细览 | 超过30w | 14 |
| 未找到标准号 | 7 |
| 重复专利树 | 1412 |
| 正常导出专利数 | 244036 |
| 正常导出专利国别统计 | country | total |
| WO | 53970 |
| USRE | 1 |
| US | 82191 |
| NO | 1 |
| NL | 4 |
| LU | 3 |
| KR | 18204 |
| JP | 66586 |
| GB | 114 |
| FR | 103 |
| EP | 24087 |
| DE | 173 |
| CH | 3 |
| CA | 7 |
| BE | 6 |
| AU | 15 |
| AT | 4 |
| 正常导出序列数量 | country | total |
| AT | 23 |
| AU | 134 |
| BE | 35 |
| CA | 34 |
| CH | 44 |
| DE | 1962 |
| EP | 4752696 |
| FR | 1123 |
| GB | 1180 |
| JP | 7182688 |
| KR | 779622 |
| LU | 8 |
| NL | 18 |
| NO | 8 |
| US | 11278436 |
| WO | 6728982 |

## 20151013V1.4

|  |  |  |  |
| --- | --- | --- | --- |
| 集合 | 总数据 | 未解析出 原始号 | 语法 |
| SourceEMBL | 39870890 | 19 | db.sourceEMBL.find({"PNO":null}).count() |
| SourceFasta | 39712183 |  |  |
| 集合 | 总数据 | 无标准号 | 语法 |
| ExchangePNO | 240722 | 7 | db.ExchangePNO.find({"StdInfo":null}).count() |

原始专利国别统计：

|  |  |  |
| --- | --- | --- |
|  | **country** | **count** |
| 1 | "AT" | 4 |
| 2 | "AU" | 15 |
| 3 | "BE" | 6 |
| 4 | "CA" | 7 |
| 5 | "CH" | 3 |
| 6 | "DE" | 173 |
| 7 | "EP" | 23113 |
| 8 | "FR" | 103 |
| 9 | "GB" | 114 |
| 10 | "JP" | 63062 |
| 11 | "KR" | 16924 |
| 12 | "LU" | 3 |
| 13 | "NL" | 4 |
| 14 | "NO" | 1 |
| 15 | "US" | 75020 |
| 16 | "USRE" | 1 |
| 17 | "WO" | 52504 |

## 数据转换情况

## 导出数据情况

EMBL导出情况统计：

|  |  |  |
| --- | --- | --- |
| 上一版本 | 专利数 | EMBL数 |
| 原始总量 | 231057 | 37030433 |
| 超过30w | 12 | 8684327 |
| 未解析专利号 |  | 19 |
| 未找到标准号 | 7 | 443 |
| 重复 | 1405 | 28345644 |
| 正确导出 | 229633 |
| 合计 | 231057 | 37030433 |
| 20151013V1.4 |  |  |
|  | 专利数 | EMBL数 |
| 原始总量 | 240722 | 39870890 |
| 超过30w | 14 | 10110583 |
| EMBL记录缺失 | 3 | 19 |
| 未找到标准号 | 7 | 443 |
| 重复 | 1407 | 29759845 |
| 正确导出 | 239291 |
| 合计 | 240722 | 39870890 |
| 20151231V1.5 |  |  |
|  | 专利数 | EMBL数 |
| 原始总量 | 245472 | 39870890 |
| 超过30w | 14 | 10110583 |
| EMBL记录缺失 | 3 | 19 |
| 未找到标准号 | 7 | 443 |
| 重复 | 1407 | 29759845 |
| 正确导出 | 244036 |  |
| 合计 | 240722 | 39870890 |

// EMBL未解析专利号

{PNO:null}

// EMBL未找到标准号

{PNO:{$in:["JP03074534T","EP2607483A1","EP2589606A1","EP2543365A1","EP2639585A1","US7138514A","US7135460A"]}}

导出序列数量 29759845

分国别统计：

|  |  |  |
| --- | --- | --- |
|  | **country** | **count** |
| 1 | "AT" | 23 |
| 2 | "AU" | 134 |
| 3 | "BE" | 35 |
| 4 | "CA" | 34 |
| 5 | "CH" | 44 |
| 6 | "DE" | 1962 |
| 7 | "EP" | 4612494 |
| 8 | "FR" | 1123 |
| 9 | "GB" | 1180 |
| 10 | "JP" | 6534407 |
| 11 | "KR" | 655575 |
| 12 | "LU" | 8 |
| 13 | "NL" | 18 |
| 14 | "NO" | 8 |
| 15 | "US" | 11278276 |
| 16 | "WO" | 6674524 |

导出专利数量 239291

导出专利起止日期 19160321 - 20150528

分国别情况：

|  |  |  |
| --- | --- | --- |
|  | **country** | **count** |
| 1 | AT | 4 |
| 2 | AU | 15 |
| 3 | BE | 6 |
| 4 | CA | 7 |
| 5 | CH | 3 |
| 6 | DE | 173 |
| 7 | EP | 23048 |
| 8 | FR | 103 |
| 9 | GB | 114 |
| 10 | JP | 64622 |
| 11 | KR | 16923 |
| 12 | LU | 3 |
| 13 | NL | 4 |
| 14 | NO | 1 |
| 15 | US | 81179 |
| 16 | WO | 53086 |

**其他：、**

**专利重复（patentrepeat）：1407**

|  |  |  |
| --- | --- | --- |
|  | **country** | **count** |
| 1 | "KR" | 1 |
| 2 | "EP" | 393 |
| 3 | "WO" | 4 |
| 4 | "US" | 1009 |

专利重复专利号样例：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **PNO** | **kind** | **PNS** | **Date** |
| 1 | EP0116778A | A | EP0116778A1 |  |
| 2 | EP0116778A1 | A1 | EP0116778A1 |  |
| 3 | EP0117063A | A | EP0117063A1 |  |
| 4 | EP0117063A1 | A1 | EP0117063A1 |  |
| 5 | EP0117767A | A | EP0117767A1 |  |
| 6 | EP0117767A1 | A1 | EP0117767A1 |  |
| 7 | EP0119031A | A | EP0119031A2 |  |
| 8 | EP0119031A2 | A2 | EP0119031A2 |  |
| 9 | KR1019940700536A | A | KR20050084550A | 20050826 |
| 10 | KR1020057004382A | A | KR20050084550A | 20050314 |
| 11 | US6245337 |  | US6245337B1 |  |
| 12 | US6245337A | A | US6245337B1 | 2001061 |
| 13 | WO0055330 |  | "WO0055330A1" |  |
| 14 | WO0055330A1 |  | "WO0055330A1" | 20000921 |
| 15 | WO8600078 |  | "WO8600078A1" |  |
| 16 | WO8600078A1 | A1 | "WO8600078A1" | 19860103 |
| 17 | WO9008187A | A | "WO9008187A1" | 19900726 |
| 18 | WO9008187A1 | A1 | "WO9008187A1" | 19900726 |

专利重复对应序列列表样例：

{PNO:{$in:["EP0116778A","EP0116778A1","EP0117063A","EP0117063A1","EP0117767A","EP0117767A1","EP0119031A","EP0119031A2","KR1019940700536A","KR1020057004382A","US6245337","US6245337A","WO0055330","WO0055330A1","WO8600078","WO8600078A1","WO9008187A","WO9008187A1"]}}

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **\_id** | **PNO** | **PNODT** | **FilePath** |
| 1 | I04016 | EP0116778A | 19840829 | rel\_pat\_unc\_13\_r122.dat |
| 2 | I04015 | EP0116778A | 19840829 | rel\_pat\_unc\_13\_r122.dat |
| 3 | I04014 | EP0116778A | 19840829 | rel\_pat\_unc\_13\_r122.dat |
| 4 | I04013 | EP0116778A | 19840829 | rel\_pat\_unc\_13\_r122.dat |
| 5 | I04012 | EP0116778A | 19840829 | rel\_pat\_unc\_13\_r122.dat |
| 6 | I04011 | EP0116778A | 19840829 | rel\_pat\_unc\_13\_r122.dat |
| 7 | I04009 | EP0116778A | 19840829 | rel\_pat\_unc\_13\_r122.dat |
| 8 | I04008 | EP0116778A | 19840829 | rel\_pat\_unc\_13\_r122.dat |
| 9 | I04007 | EP0116778A | 19840829 | rel\_pat\_unc\_13\_r122.dat |
| 10 | I04006 | EP0116778A | 19840829 | rel\_pat\_unc\_13\_r122.dat |
| 11 | I04005 | EP0116778A | 19840829 | rel\_pat\_unc\_13\_r122.dat |
| 12 | AAA56463 | EP0116778A1 | 19840829 | uspto\_prt.dat |
| 13 | I04044 | EP0117063A | 19840829 | rel\_pat\_unc\_13\_r122.dat |
| 14 | I04039 | EP0117063A | 19840829 | rel\_pat\_unc\_13\_r122.dat |
| 15 | AAA56462 | EP0117063A1 | 19840829 | uspto\_prt.dat |
| 16 | AAA56461 | EP0117063A1 | 19840829 | uspto\_prt.dat |
| 17 | AAA56460 | EP0117063A1 | 19840829 | uspto\_prt.dat |
| 18 | AAA56459 | EP0117063A1 | 19840829 | uspto\_prt.dat |
| 19 | AAA56458 | EP0117063A1 | 19840829 | uspto\_prt.dat |
| 20 | I04084 | EP0117767A | 19840905 | rel\_pat\_unc\_13\_r122.dat |
| 21 | I04083 | EP0117767A | 19840905 | rel\_pat\_unc\_13\_r122.dat |
| 22 | I04076 | EP0117767A | 19840905 | rel\_pat\_unc\_13\_r122.dat |
| 23 | AAA56457 | EP0117767A1 | 19840905 | uspto\_prt.dat |
| 24 | AAA56456 | EP0117767A1 | 19840905 | uspto\_prt.dat |
| 25 | I07861 | EP0119031A | 19840919 | rel\_pat\_unc\_13\_r122.dat |
| 26 | I07860 | EP0119031A | 19840919 | rel\_pat\_unc\_13\_r122.dat |
| 27 | I07859 | EP0119031A | 19840919 | rel\_pat\_unc\_13\_r122.dat |
| 28 | AAA55497 | EP0119031A2 | 19840919 | uspto\_prt.dat |
| 29 | AAA55496 | EP0119031A2 | 19840919 | uspto\_prt.dat |
| 30 | AAA55495 | EP0119031A2 | 19840919 | uspto\_prt.dat |
| 31 | DI507929 | KR1019940700536A | 19940222 | kipo\_prt.dat |
| 32 | DI105721 | KR1020057004382A | 20050314 | rel\_pat\_hum\_02\_r122.dat |
| 33 | DI105486 | KR1020057004382A | 20050314 | rel\_pat\_hum\_02\_r122.dat |
| 34 | AR157320 | US6245337 |  | rel\_pat\_unc\_01\_r122.dat |
| 35 | AAE73921 | US6245337A | 20010612 | uspto\_prt.dat |
| 36 | AAE73920 | US6245337A | 20010612 | uspto\_prt.dat |
| 37 | AAE73919 | US6245337A | 20010612 | uspto\_prt.dat |
| 38 | AAE73918 | US6245337A | 20010612 | uspto\_prt.dat |
| 39 | AAE73917 | US6245337A | 20010612 | uspto\_prt.dat |
| 40 | AAE73916 | US6245337A | 20010612 | uspto\_prt.dat |
| 41 | AAE73915 | US6245337A | 20010612 | uspto\_prt.dat |
| 42 | AAE73914 | US6245337A | 20010612 | uspto\_prt.dat |
| 43 | AX035537 | WO0055330 |  | rel\_pat\_inv\_01\_r122.dat |
| 44 | AX035538 | WO0055330A1 | 20000921 | rel\_pat\_inv\_01\_r122.dat |
| 45 | AX035536 | WO0055330A1 | 20000921 | rel\_pat\_inv\_01\_r122.dat |
| 46 | AX035535 | WO0055330A1 | 20000921 | rel\_pat\_inv\_01\_r122.dat |
| 47 | AX035534 | WO0055330A1 | 20000921 | rel\_pat\_inv\_01\_r122.dat |
| 48 | AX035533 | WO0055330A1 | 20000921 | rel\_pat\_inv\_01\_r122.dat |
| 49 | AX035532 | WO0055330A1 | 20000921 | rel\_pat\_inv\_01\_r122.dat |
| 50 | AX035531 | WO0055330A1 | 20000921 | rel\_pat\_inv\_01\_r122.dat |
| 51 | AX035530 | WO0055330A1 | 20000921 | rel\_pat\_inv\_01\_r122.dat |
| 52 | AX035529 | WO0055330A1 | 20000921 | rel\_pat\_inv\_01\_r122.dat |
| 53 | AX035528 | WO0055330A1 | 20000921 | rel\_pat\_inv\_01\_r122.dat |
| 54 | AX035527 | WO0055330A1 | 20000921 | rel\_pat\_inv\_01\_r122.dat |
| 55 | AX035526 | WO0055330A1 | 20000921 | rel\_pat\_inv\_01\_r122.dat |
| 56 | AX035525 | WO0055330A1 | 20000921 | rel\_pat\_inv\_01\_r122.dat |
| 57 | AX035546 | WO0055330A1 | 20000921 | epo\_prt.dat |
| 58 | AX035545 | WO0055330A1 | 20000921 | epo\_prt.dat |
| 59 | AX035544 | WO0055330A1 | 20000921 | epo\_prt.dat |
| 60 | AX035543 | WO0055330A1 | 20000921 | epo\_prt.dat |
| 61 | AX035542 | WO0055330A1 | 20000921 | epo\_prt.dat |
| 62 | AX035541 | WO0055330A1 | 20000921 | epo\_prt.dat |
| 63 | AX035540 | WO0055330A1 | 20000921 | epo\_prt.dat |
| 64 | AX035539 | WO0055330A1 | 20000921 | epo\_prt.dat |
| 65 | A01041 | WO8600078 |  | epo\_prt.dat |
| 66 | A01040 | WO8600078A1 | 19860103 | epo\_prt.dat |
| 67 | A01039 | WO8600078A1 | 19860103 | epo\_prt.dat |
| 68 | AAA55025 | WO9008187A | 19900726 | uspto\_prt.dat |
| 69 | AAA55024 | WO9008187A | 19900726 | uspto\_prt.dat |
| 70 | AAA55023 | WO9008187A | 19900726 | uspto\_prt.dat |
| 71 | A00768 | WO9008187A1 | 19900726 | epo\_prt.dat |
| 72 | I06836 | WO9008187A1 | 19900726 | rel\_pat\_unc\_13\_r122.dat |
| 73 | I06833 | WO9008187A1 | 19900726 | rel\_pat\_unc\_13\_r122.dat |

**单专利Embl记录大于28w (ExchangePNO.state==3）**

|  |  |  |
| --- | --- | --- |
|  | **country** | **count** |
| 1 | "EP" | 1 |
| 2 | "JP" | 4 |
| 3 | "US" | 2 |
| 4 | "WO" | 5 |

数据：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **PNO** | **kind** | **PNS** | **序列数量** |
| 1 | EP2213738A2 | A2 | EP2213738A2 | 1590133 |
| 2 | JP2006507841A | A | JP2006507841A | 1591911 |
| 3 | JP2007536925A | A | JP2007536925A | 760616 |
| 4 | JP2008500039A | A | JP2008500039A | 547372 |
| 5 | JP2008518610A | A | JP2008518610A | 483996 |
| 6 | US7374927A | A | US7374927B2 | 673904 |
| 7 | US7618814A | A | US7618814B2 | 548156 |
| 8 | WO2005116204A | A | WO2005116204A1 | 817651 |
| 9 | WO2005116250A2 | A2 | WO2005116250A2 | 547372 |
| 10 | WO2005116265A2 | A2 | WO2005116265A2 | 362830 |
| 11 | WO2007106407A2 | A2 | WO2007106407A2 | 478375 |
| 12 | WO2005014857A2 | A2 | WO2005014857A2 | 282011 |

对应序列数量：{PNO:{$in:[" EP2213738A2","JP2006507841A","JP2007536925A","JP2008500039A","JP2008518610A","US7374927A","US7618814A","WO2005116204A","WO2005116250A2","WO2005116265A2","WO2007106407A2","WO2005014857A2"]}

如：超过MongoDB最大字段内容限制：WO2005014857A2 （序列数量282021

**未匹配标准号（ExchangePNO.StdInfo==null）：**

|  |  |  |
| --- | --- | --- |
|  | **country** | **count** |
| 1 | "JP" | 1 |
| 2 | "EP" | 4 |
| 3 | "US" | 2 |

数据：

|  |  |  |  |
| --- | --- | --- | --- |
| EMBL专利号 | EMBL日期 | PNS | 序列号 |
| JP03074534T | 20010119 |  | BD004377 |
| EP2607483A1 | 20130626 |  | JB754055 |
| EP2589606A1 | 20130508 |  | JB420465 |
| EP2543365A1 | 20130109 |  | JB343810 |
| EP2639585A1 | 20130918 |  | JC011835 |
| US7138514A | 20061121 |  | EA028895 |
| US7135460A | 20061114 |  | ABL32763 |

**未解析标准号：**

|  |
| --- |
| **\_id** |
| "A06505" |
| "A34685" |
| "A34668" |
| "A34667" |
| "A34666" |
| "A34665" |
| "A34642" |
| "A34641" |
| "A21495" |
| "A20460" |
| "A14664" |
| "A14659" |
| "A14254" |
| "A14131" |
| "A13360" |
| "A12877" |